Towards a resource for multilingual lexicons: an MT assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation

Han, L.; Mohamed, N. H.; Rassem, M.; Jones, G. J.F.; Smeaton, A. F.; Nenadic, G.

Noname manuscript No.
(will be inserted by the editor)

Towards a resource for multilingual lexicons: an MT assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation

Lifeng Han · Najet Hadj Mohamed · Malak Rassem · Gareth J.F. Jones · Alan F. Smeaton · Goran Nenadic

Received: 12 Aug 2021 / Accepted: 25 October 2025

Abstract In this work, we introduce the construction of a machine translation (MT) assisted and human-in-the-loop multilingual parallel corpus with annotations of multi-word expressions (MWEs), named AlphaMWE. The MWEs include verbal MWEs (vMWEs) defined in the PARSEME shared task that have a verb as the head of the studied terms. The annotated vMWEs are also bilingually and multilingually aligned manually. The languages covered include Arabic, Chinese, English, German, Italian, and Polish, of which, the Arabic corpus includes both standard and dialectal variations from Egypt and Tunisia. Our original English corpus is extracted from the PARSEME shared task in 2018. We performed machine translation of this source corpus followed by human post-editing and annotation of target MWEs. Strict quality control

LH
University of Manchester, Manchester M13 9PL, The United Kingdom
Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, NL
Leiden University Medical Center, Leiden University, Leiden, NL
E-mail: l.han@lumc.nl, lifeng.han@manchester.ac.uk

NM
University of Tours, France and Arabic Natural Language Processing Research Group, University of Sfax, Tunisia

MR
Institute for Natural Language Processing (IMS), University of Stuttgart, Germany

GJ
ADAPT Research Centre, Dublin City University, Dublin 9, Ireland
E-mail: gareth.jones@adaptcentre.ie

AS
Insight Centre for Data Analytics, Dublin City University, Dublin 9, Ireland
E-mail: alan.smeaton@dcu.ie

GN
University of Manchester, Manchester M13 9PL, The United Kingdom
E-mail: g.nenadic@manchester.ac.uk

was applied for error limitation, i.e., each MT output sentence received first manual post-editing and annotation plus a second manual quality rechecking till annotators' consensus is reached. One of our findings during corpora preparation is that accurate translation of MWEs presents challenges to MT systems, as reflected by the outcomes of human-in-the-loop metric HOPE. To facilitate further MT research, we present a categorisation of the error types encountered by MT systems in performing MWE-related translation. To acquire a broader view of MT issues, we selected four popular state-of-the-art MT systems for comparison, namely Microsoft Bing Translator, GoogleMT, Baidu Fanyi, and DeepL MT. Because of the noise removal, translation post-editing, and MWE annotation by human professionals, we believe the AlphaMWE data set will be an asset for both monolingual and cross-lingual research, such as multi-word term lexicography, MT, and information extraction.

Keywords Multi-word Expression · Parallel Corpus · Multilingual Resource Construction · Machine Translation · Translation Post-editing · Manual Annotation · Translation Evaluation

1 Introduction

[1] Multi-word Expressions (MWEs) have long been of interest to both natural language processing (NLP) researchers and linguists [28], [5], [23]. The automatic processing of MWEs has posed significant challenges for some tasks in computational linguistics (CL), such as word sense disambiguation (WSD), parsing, and (automated) translation [14], [3], [31], [16], [9]. This is caused by both the variety and the richness of MWEs as they are used in language.

Various definitions of MWEs have included both syntactic structure and semantic viewpoints from different researchers covering syntactic anomalies, non-compositionality, non-substitutability and ambiguity [5]. For instance, [1] define MWEs as "lexical items that: (i) can be decomposed into multiple lexemes; and (ii) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". The annual MWE workshop organized by SIGLEX-MWE group (Special Interest Group in Lexicon), from the Association for Computational Linguistics (ACL)[2], has focused on the interests on MWEs in different fields of NLP research. A scientific shared task on MWE discovery and identification has been held for 3 years to date (2017, 2018, and 2020)[19, 26, 20], where monolingual corpus with verbal MWEs annotated was offered for challenges across different languages, mainly European ones. Another related workshop series that had been held every two years were from EUROPHRAS " Workshop on Multi-word Units in Machine Translation and Translation Technologies" since 2013[3]. However, as noted by other NLP researchers for example in [5],

[1] Accepted manuscript by Language Resources and Evaluation Journal `http://www.elda.org/en/dissemination/jlre-language-resources-and-evaluation-journal/`

[2] `http://multi-word.sourceforge.net/PHITE.php`

[3] http://www.lexytrad.es/europhras2019/mumttt-2019-2/

there are very few bilingual or even multilingual parallel corpora with MWE annotations available for cross-lingual NLP research and for downstream applications such as machine translation (MT) [11, 40].

With regard to MWE research, verbal MWEs (vMWEs) are a mature category that have received attention from many researchers [18, 22]. Verbal MWEs have a verb as the head of the studied term and function as verbal phrases, such as "*kick* the bucket", "*cutting* capers" and "*go* to one's head". In this work, we present the construction of a multilingual corpus with vMWEs annotation, including English-Chinese, English-German, English-Polish, English-Italian, and English-Arabic language pairs, with a main focus on English-Chinese and English-Arabic language pairs. We started with the same source monolingual corpus in English with its vMWE tags from the shared task affiliated with the SIGLEX-MWE workshop in 2018 [38, 26]. Several state-of-the-art (SOTA) MT models were used to perform an automated translation, and then human post-editing and annotation for the target languages was conducted with cross-validation to ensure the quality, i.e., with each sentence receiving post-editing and rechecking by at least two people. For dialectal Arabic, we carried out the translation from scratch without MT in the loop because the current state-of-the-art systems don't have such options and it is easier to translate manually than post-editing.

In order to get a deeper insight into the difficulties of processing MWEs we carried out a categorisation of the errors made by MT models when processing MWEs, especially using English-Chinese language pairs. We also produced quantitative analysis scores using the human post-editing metric HOPE on the English-Arabic language. From these, we conclude that current state-of-the-art MT models are far from reaching parity with humans in terms of translation performance, especially on idiomatic MWEs, and even for sentence level translation, although some researchers sometimes claim otherwise [39, 10].

The rest of this paper is organised as follows. In the next section (2), we present related work and then detail the corpus preparation stages including selection of MT models and the resulting AlphaMWE (Section 3). We then look at various issues that MT has with MWEs (Section 4), followed by a quantitative evaluation section using post-editing metric HOPE [6] on English-Arabic language pair as a case study (Section 5). Subsequently, we present a broader discussion section (6). This analysis and discussion, along with the public release of the corpora as a resource for the community, is the main contribution of the paper [4]. Finally, we conclude the paper with a plan for future work in Section 7. This is an extended work of our earlier MWE-workshop paper [8], where we add two more language pairs English-Italian and English-Arabic [21] for the corpora, add Polish and German sub-sections on error analysis from MT systems, quantitative evaluation, extensive discussion and analysis sections on MWEs related MT issues.

[4] We adopt the same license as the original PARSEME English dataset, i.e. CC-BY-SA 4.0

2 Related Work

There is a number of existing studies that focus on the creation of *monolingual* corpora with vMWE annotations, such as the PARSEME shared task corpora [29, 26]. The 2020 edition of this task covers 14 languages including Chinese, Hindi, and Turkish as non-European languages. [36] prepared a manual annotation of Hungarian corpus with Light Verb Constructions (LVCs) and indicated their usability for machine translation and information extraction. Some work from monolingual English corpora includes the "MWE-aware English Dependency Corpus" from the Linguistic Data Consortium (LDC2017T01) [12] that covers *compound words* used to train parsing models, and Wiki50 by [37] which consists of 50 Wikipedia articles (4,350 sentences) with the annotation of MWEs (compound, verb-particle constructions, idiom, Light verb construction, multi-word verbs) and named entities (PER, LOC, ORG and MISC). Also related to this are English MWEs from "web reviews data" by [30] that covers *noun, verb* and *preposition super-senses* and English verbal MWEs from [38] and [13] that covers PARSEME shared task defined vMWE categories. However, all these works were performed in monolingual settings, independently by different language speakers without any bilingual alignment. These corpora are helpful for monolingual MWE research such as *discovery* or *identification*, however, it would be difficult to use these corpora for bilingual or multilingual research such as MT or cross-lingual information extraction.

The work most related to ours is the one by [35], who created an English-Hungarian parallel corpus with annotations for light verb constructions (LVCs). As many as 703 LVCs for Hungarian and 727 for English were annotated in this work, and a comparison between English and Hungarian data was carried out. LVCs often have a noun and a stretched verb, where the noun keeps its literal senses while the verb "loses its original sense to some extent" [35], for instance, "take a bite out of..." which is similar to the verb "bite", "did cleaning", and "taken into consideration", etc. However, the work did not cover other types of vMWEs, for instance inherently adpositional verbs, verbal idioms, or verb-particle constructions, and it was not extended to any other language pairs. In our work, we annotate in a multilingual setting including distant languages such as German (Germanic), Polish (Slavic), Italian (Romance), Chinese, and Arabic, in addition to the extension of vMWE categories. In other recent work [9], we performed an automatic construction of bilingual MWE terms based on a parallel corpus, in this case, English-Chinese and English-German. We first conducted automated extraction of monolingual MWEs based on part-of-speech (POS) patterns and then aligned the monolingual MWEs into bilingual terms based on statistical lexical translation probability. However, due to the automated procedure, the extracted bilingual "MWE terms" contain not only MWEs but also normal phrases. Part of the reason for this is due to the POS pattern design which is a challenging task for each language and needs to be further refined [31], [27], [9].

In summary, in this work we aim at creating a multilingual parallel corpus that contains broader vMWE categories introduced by PARSEME shared task

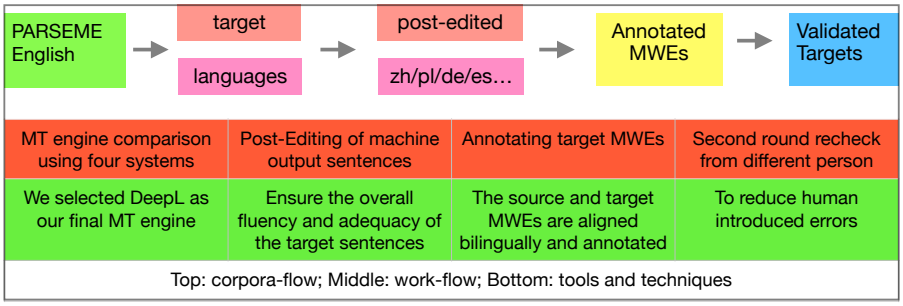


Fig. 1 Workflow to prepare the AlphaMWE corpus

[19, 26, 20] but in a parallel and multilingual setting. This will benefit both the MWE identification task by extending the languages covered and down-stream multilingual applications especially MT since there is still a lack of such data on MWEs-aware test sets for MT system testing, especially high-quality and manually annotated ones [35, 31, 9].

The potential applications of AlphaMWE include its integration into MT systems similar to [33], who applied bilingual MWE and named entities corpora into English-Hindi (EN-HI) and Hindi-English (HI-EN) MT, and [40, 9], who performed MWE aware NMT. Task-oriented MT evaluation is still in need of creating corresponding data sets [15], and AlphaMWE may be used for such a purpose, namely to test an MT model's ability for translating idiomatic MWEs as one of the main indicators to reach human parity. For instance, the very recent work by [2] created an English-French corpus for MT evaluation on a dialogue task; [17] emphasised idioms as challenges for English↔German MT systems in WMT2021 shared task; our own work on human-centric and task-oriented metric HOPE [6] using professional post-editing annotation for MT evaluation (MTE) uses optimised error categories which can be deployed using the AlphaMWE corpus for testing MT engines and we will give an example testing using English-Arabic corpus. AlphaMWE would also be an asset for MWE identification and information extraction in a multilingual scenario, e.g. [37] conducted identification of noun compounds and named entities with the monolingual MWE corpus they prepared.

3 Corpus Construction: AlphaMWE

We now describe our corpus preparation method, the selection of the MT models used in our investigation, and the resulting open-source corpora AlphaMWE.

3.1 *Corpus Preparation*

To construct a well-aligned multilingual parallel corpus, our approach is to take a monolingual corpus from the PARSEME *vMWE discovery and identification shared task* as our root corpus. Our rationale is that this shared task is well established and its process of tagging and categorisation is clear. Furthermore, as we plan to extend the MWE categories in the future e.g., not only the verbal ones, we enrich the PARSEME shared task corpus with potential for other downstream research and applications, including bilingual and multilingual NLP models such as MT. The English corpus [38] we used from the PARSEME shared task follows the annotation guidelines having a broad range of vMWE categories tagged. These include inherently adpositional verbs, light verb constructions, multi-verb constructions, verbal idioms, and verb-particle constructions. The English corpus contains sentences from several different domains such as news, literature, and IT documents, which were observed during our human post-editing procedure. For the IT document domain, vMWEs are usually easier or more straightforward to translate, with a high chance of repetition, e.g. "apply filter" and "based on", which are mostly translated correctly by the MT engines. For the literature annotations, the vMWEs include richer samples with many idiomatic or metaphorical expressions, such as "cutting capers" and "gone slightly to someone's head" that cause MT issues.

Fig. 1 shows our workflow. We first use MT models to perform automated translation for the target language direction and then carry out human post-editing of the output hypotheses with annotation of the corresponding target-side vMWEs which are aligned with the source English ones. In order to choose which MT system to use for producing the translations, we perform manual inspection on 10 (randomly picked) outputs from these systems: DeepL, GoogleMT, Bing Translator, and Baidu Translator. The rationale for the MT+PE design is that firstly the state-of-the-art NMT engines can produce relatively good results even though often with a certain degree of mistakes, which already saves a lot of time in comparison to human translation from scratch; secondly, one of the motivations of our corpus construction is to apply it to a multilingual NLP setting especially MT, and we are curious to examine how good (or bad) the MT performance is in term of MWEs. Finally, to avoid human-introduced errors, we apply a cross-validation strategy, where each sentence receives at least a second person's quality checking after the first post-editing. Tagging errors are more likely to occur if only one human has seen each sentence and we discuss some error samples from English source corpus in later sections.

The instructions given to the post-editors and annotators include i) edit any mistranslations that occurred with a meaning-equivalency aspect from source English to the target language; ii) wherever possible, try to use corresponding MWEs in the target language for the terms that are highlighted as MWEs in the source English; iii) offer alternative translations regarding some phrase/expression translations wherever suitable. iv) try to keep the register-level the same from the source language to the target language, e.g. 'thx' *vs*

thanks, a word of lower register that occurs in some social media data; v) wherever an ambiguous translation occurs, the background context information shall be referred to, which is available to all post-editors. Overall the post-editing and annotation tasks were evenly shared among workers except for English-Chinese and English-Italian where more experienced annotators have carried out larger amount of annotations than less experienced annotators.

The background of the post-editing and annotation workers are a) English-German: two research students who are German native speakers including one master's student and one Ph.D. candidate with backgrounds in linguistics and informatics (both graduated now); b) English-Polish: two Ph.D. candidates who are Polish native speakers with a background of linguistics and semantics ('semantics' one graduated); c) English-Chinese: four Chinese Mandarin native speakers including two Ph.D. candidates from applied linguistics and NLP ('NLP' one graduated), and two full-time faculty from the applied linguistics and child language learning areas (both hold a Ph.D. degree); d) English-Italian: two native Italian speakers one of whom is a postdoctoral researcher in computer science and informatics, and the other works in Ireland for 5 years in customer service sector as a fluent English speaker; e) English-Arabic: one Ph.D. candidate in NLP focusing on MWEs and one Masters student in Translation Studies.

3.2 *MT System Selection*

We tested a number of example sentences (approximately 10) from the English test set to compare state-of-the-art MT from Microsoft Bing [4], GoogleMT [34], Baidu Fanyi [32], and DeepL[5], as in Fig. 2.[6]. We illustrate the comparative performances with two representative example translations, which reflect the quality differences between these engines on the corpus we are using. As a first example sentence, GoogleMT and Bing Translator have very similar outputs, where the MT engines have a tendency to produce as much information as possible, but make the sentences redundant or awkward to read, such as the phrase "验证... 是否验证了 (yàn zhèng ... Shì fǒu yàn zhèng le)" where they use a repeated word "验证" (yàn zhèng, *verify*). Although the DeepL Translator does not produce a perfect translation since it drops the source word "validated" which should be translated as "有效性 (yǒu xiào xìng)"(as one candidate translation), the overall output is fluent and the source sentence meaning is mostly preserved. Baidu translator yields the worst output in this example. It produces some words that were not in the source sentence (或者, huò zhě, *or*), loses some important terms' translation from the source sentence ("SQL Server", the subject of the sentence), and the reordering of the sentence fails resulting in an incorrect meaning ("在没有密码的情况下, zài méi yǒu mì

[5] https://www.deepl.com/en/translator (All testing was performed in 2020/07 from 4 MT models)

[6] All testing was performed in 2020/07

mǎ de qíng kuàng xià" is moved from the end of the sentence to the front and made as a condition). So, in this case, DeepL performed best.

In a second example sentence, GoogleMT confused the original term TSQL as SQL. Bing MT had a similar issue with the last example, i.e. it produced redundant information "有关 (yǒu guān)" (*about/on*). In addition, it concatenated the website address and normal phrase "了解有关 (liǎo jiě yǒu guān)" together with a hyperlink. GoogleMT and Bing both translated half of the source term/MWE "Microsoft Developer Network Web" as "Microsoft 开发人员网络网站" (kāi fā rén yuán wǎng luò wǎng zhàn) where they kept "Microsoft" but translated "Developer Network Web". Although this is a reasonable output since Microsoft is a general popular named entity while "Developer Network Web" consists of common words, we interpret "Microsoft Developer Network Web" as a named entity/MWE in the source sentence that consists of all capitalised words which would be better translated overall as "微软开发人员网络网站 (wēi ruǎn kāi fā rén yuán wǎng luò wǎng zhàn)" or be kept as the original capitalised words as a foreign term in the output, which is how DeepL outputs this expression. However, Baidu totally drops out this MWE translation and another word translation is not accurate, "more" into 详细 (xiáng xì). These samples illustrate why we chose to use DeepL as the provider of our MT translations. Similar comparisons were carried out on English-German and English-Polish which also favoured DeepL.

3.3 *Outcome: AlphaMWE Corpus*

Regarding the size of the corpus, we extracted all 750 English sentences which have vMWE tags included. The target languages are Arabic, Chinese, German, Polish, and Italian with sample sentences in Fig. 3. Our very recent work on confidence sample size estimation for MT evaluation using Monte Carlo sampling simulation indicates that the test set with 200+ sentences can reflect MT quality in a sufficient matter [7]. This means that our bilingual/multilingual corpora with 750 sentences are suitable for shared task usage, specifically to a linguistic phenomenon.

There are several situations and decisions that are worth noting during the post-editing (PE) stage: a) when the original English vMWEs are translated into a general phrase in the target language but not choosing a sequence of MWEs, we tried to offer two different references, with one of them being revised in a vMWE/MWE presentation in the target. For instance "(the beer had) slightly gone to his head" can be translated as "微微让他上了头 (wéi wéi ràng tā shàng le tóu, he got a little drunk)" or "他微微醉了 (tā wéi wéi zuì le, he got a little drunk)" of which the first term "上了头 (shàng le tóu, literally 'up to the head')" is an idiomatic MWE describing the concept "醉了 (zuì le, literally "drunk")"; b) when the original English sentence terms were correctly translated into the target language but in a different register, e.g. the source language has a low register ('thx', for instance), we asked the post-editors to offer two reference sentences, with one of them using the same low register

Two sample sentences' MT outputs comparison from head of test file	
Source	# text = SQL Server verifies that the account name and password were validated when the user logged on to the system and grants access to the database, without requiring a separate logon name or password.
DeepL	# text = SQL Server 会在用户登录系统时验证账户名和密码，并授予对数据库的访问权限，而不要求单独的登录名或密码。 *en: text = SQL verifies the account name and password when users are logging into the system, and grants the access to the database, without requiring a separate logon name or password.*
Google	# text = SQL Server验证用户登录系统时是否验证了帐户名和密码，并授予对数据库的访问权限，而无需单独的登录名或密码。(yàn zhèng … shì fǒu yàn zhèng le) *en: text = "SQL server" verifies if users verified the account name and password when they are logging into the system, and grants the access to the database, without the need of a separate logon name or password.*
Bing	[text] SQL Server 验证用户登录到系统时是否验证了帐户名称和密码，并授予对数据库的访问权限，而无需单独的登录名称或密码。(yàn zhèng … shì fǒu yàn zhèng le) *en: [text] "SQL Server" verifies if users verified the account name and password when they logged into the system, and grants the access to the database, without the need of a separate logon name or password.*
Baidu	#或者，在没有密码的情况下验证用户名和用户名是否被登录到数据库中，并且系统是否授予登录权限。(huò zhě, zài méi yǒu mì mǎ de qíng kuàng xià) *en: or, without the password, verify account name and if account name is logged into the database, and if the system grants access to login.*
Ref.	# 文本 = SQL Server 会在用户登录系统时验证账户名和密码的有效性，并授予对数据库的访问权限，而不要求单独的登录名或密码。
Source	# text = See the http://officeupdate.microsoft.com/, Microsoft Developer Network Web site for more information on TSQL.
DeepL	# text = 有关 TSQL 的更多信息，请参见 http://officeupdate.microsoft.com/，Microsoft Developer Network Web 站点。 *en: text = For more information about "TSQL", please see "http://officeupdate.microsoft.com/", the website "Microsoft Developer Network Web".*
Google	# text =有关Microsoft SQL的更多信息，请参见http://officeupdate.microsoft.com/，Microsoft开发人员网络网站。 *en: text = For more information about "Microsoft SQL", please see " http://officeupdate.microsoft.com/", "Microsoft" developers' network web.*
Bing	[文本] 有关 TSQL http://officeupdate.microsoft.com/了解有关 TSQL 的相关信息，请参阅 Microsoft 开发人员网络网站。 *en: [text] For information about related information of "TSQL http://officeupdate.microsoft.com/了解有关 TSQL", please see "Microsoft" developers' network web.*
Baidu	#text=参见http://officeupdate.microsoft.com/，以获取有关TSQL的详细信息。 *en: text=see http://officeupdate.microsoft.com/ to get detailed information about "TSQL".*
Ref.	# text = 有关 TSQL 的更多信息，请参见 http://officeupdate.microsoft.com/，微软开发人员网络互联网站点。
Blue: redundancy; green: adding error; pink: reordering error; yellow: dropping error. *en: for easier understanding*	

Fig. 2 Sample comparison of outputs from four MT models

and the other with (formal) full word spelling; c) for the situations where a single English word or normal phrase is translated into a typical vMWE in the target language, or both source and target sentences include vMWEs but the source vMWE was not annotated in the original English corpus, we made some additions to include such vMWE (pairs) into AlphaMWE; d) for some wrong/incorrect annotation in the source English corpus or some misspelling of words, we corrected them in AlphaMWE; e) we chose English as root/source corpus, since the post-editing and annotation of target languages require the human annotators to be fluent/native in both-side languages, and all annotators were fluent in English as well as being native speakers in the specific target languages respectively.

As shown in the examples (Fig. 3) from Chinese, German, Polish, and Italian, all languages are sentence-aligned and the paired vMWEs (source, target) are listed at the end of each sentence in the order in which they appear in the sentence. AlphaMWE also includes statistics of the annotated vMWEs and a multilingual vMWEs glossary. The AlphaMWE corpora were divided

Plain English Corpus	The chair was comfortable, and the beer had gone slightly to his head. I was smoking my pipe quietly by my dismantled steamer, and saw them all cutting capers in the light, with their arms lifted high, when the stout man with mustaches came *tearing down* to the river, a tin pail in his hand, assured me that everybody was 'behaving splendidly, splendidly, dipped about a quart of water and *tore back* again. (*the italic was not annotated in source English*)
English MWEs	gone (slightly) to his head, cutting capers, tearing down, tore back
Target Chinese Corpus	椅子很舒服，啤酒已经微微让他上了头。[sourceVMWE: gone (slightly) to his head][targetVMWE: (微微)让他上了头] 我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下欢呼雀跃，高举双臂，这时，那个留着胡子的大块头，手里拿着一个铁皮桶，快速来到河边，向我确保大家都“表现得很精彩，很精彩”，他浸了大约一夸脱的水，又快速回去了。[sourceVMWE: cutting capers; tearing down; tore back][targetVMWE: 欢呼雀跃; 快速到; 快速回去]
Target German Corpus	Der Stuhl war bequem, und das Bier war ihm leicht zu Kopf gestiegen. [sourceVMWE: gone (slightly) to his head] [targetVMWE: (leicht) zu Kopf gestiegen] Ich rauchte leise meine Pfeife an meinem zerlegten Dampfer und sah, wie sie alle im Licht mit hoch erhobenen Armen Luftsprünge machten, als der stämmige Mann mit Schnurrbart mit einem Blecheimer in der Hand zum Fluss hinunterkam und mir versicherte, dass sich alle "prächtig, prächtig benahmen, etwa einen Liter Wasser eintauchte und wieder zurückwankte". [sourceVMWE: cutting capers; tearing down; tore back][targetVMWE: Luftsprünge machten; hinunterkam; zurückwankte]
Target Polish Corpus	Krzesło było wygodne, a piwo lekko uderzyło mu do głowy. [sourceVMWE: gone (slightly) to his head][targetVMWE: (lekko) uderzyło mu do głowy] Cicho paliłem swoją fajkę przy zdemontowanym parowcu i widziałem, jak wszyscy pląsają w świetle, z podniesionymi wysoko ramionami, gdy twardziel z wąsami przyszedł szybkim krokiem do rzeki, blaszany wiaderko w dłoni, zapewnił mnie, że wszyscy "zachowują się wspaniale, wspaniale, nabrał około ćwiartkę wody i zawrócił szybkim krokiem". [sourceVMWE: cutting capers; tearing down; tore back][targetVMWE: pląsają; przyszedł szybkim krokiem; zawrócił szybkim krokiem]
Target Italian Corpus	La sedia era comoda, e la birra gli aveva leggermente dato alla testa. [sourceVMWE: gone (slightly) to his head] [targetVMWE: aveva (leggermente) dato alla testa] Stavo fumando tranquillamente la pipa vicino al mio piroscafo smontato, e li ho visti tutti giocare gioiosamente alla luce, con le braccia alzate, quando l'uomo robusto con i baffi è venuto giù al fiume alacremente, un secchio di latta in mano, mi ha assicurato che tutti si stavano "comportando splendidamente, splendidamente, ha preso circa un litro d'acqua ed è tornato indietro velocemente." [sourceVMWE: cutting capers; tearing down; tore back] [targetVMWE: giocare gioiosamente; venuto giù alacremente; tornato indietro velocemente]
Target Arabic Corpus	كان الكرسي مريحًا ، وكانت الجعة قد أثرت برأسه قليلاً. [sourceVMWE: gone (slightly) to his head][targetVMWE: أثرت برأسه قليلاً]. كنت أدخن غليونتي بهدوء بواسطةأداة التبخير المفككة ، ورأيتهم جميعًا يتسللون في الضوء ، وأذرعهم مرفوعة عالياً ، عندما جاء الرجل الشجاع ذو الشوارب ينزف إلى النهر ، ودلو من الصفيح في يده ، أكد لي ذلك كان الجميع `` يتصرفون بشكل رائع ، بشكل رائع ، غمسوا حوالي لتر من الماء ومزقوا مرة أخرى. [sourceVMWE: cutting capers][targetVMWE: يتسللون]
AlphaMWE corpora examples from multilingual parallel files. “cutting capers” was annotated as VID type of MWEs, while “tearing down” and “tore back” were not annotated in the source English corpus. We added them into AlphaMWE multilingual corpora since they do cause translation errors for most state-of-the-art MT models. The bilingual MWEs are aligned with their appearance order from sentence inside the afterwards attached bracket-pairs.	

Fig. 3 AlphaMWE corpora samples with two sentences (English, Chinese, German, Polish, Italian, Arabic)

evenly into five portions which were designed in the post-editing and annotation stage. The sentences were grouped in the order that they were extracted from the original English source larger corpus with context information. As a result, it is convenient for researchers to use them for testing NLP models, choosing any subset portion or combination, or cross-validation usage.

4 MT issues with MWEs

As mentioned in the previous section, one of the motivations for creating the AlphaMWE multilingual corpus with MWEs is to apply it to multilingual NLP applications such as MT and examine how MT performs in an MWE-related context. In this section, we present an analysis of the quality of various MT systems when required to translate MWEs or MWE-related contexts. We performed the categorisation of errors during the early post-editing stage,

Source	At the corner of 72nd Street and Madison Avenue, he waved down a cab.
DeepL	在72街和麦迪逊大道的拐角处，他向一辆出租车招手。 Zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù, tā xiàng yī liàng chū zū chē zhāo shǒu.
Bing	在72街和麦迪逊大道的拐角处，他挥手示意一辆出租车。 zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù , tā huī shǒu shì yì yí liàng chū zū chē.
Google	在第72街和麦迪逊大街的拐角处，他挥舞着出租车。 Zài dì 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā huī wǔ zhe chū zū chē.
Baidu	在72街和麦迪逊大街的拐角处，他挥手叫了一辆出租车。 zài 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā huī shǒu jiào le yí liàng chū zū chē.
Ref.	在72街和麦迪逊大道的拐角处，他招手示停了一辆出租车。 Zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù, tā zhāo shǒu shì tíng le yī liàng chū zū chē.

Fig. 4 MT issues with MWEs: common sense (Pinyin is offered by GoogleMT with post-editing.)

and this categorisation was presented to all the post-editing and annotation workers during frequent communications across the editing phase.

In this paper, we mainly focus on the English→Chinese language pair. We also highlight some issues on English→German, English→Polish, and English→Arabic in different sections, but leave a detailed analysis of other language pairs for future work. Some extra examples of English-Polish MT issues related to MWEs will be presented in the Appendix section.

4.1 *English-to-Chinese*

When MT produces incorrect or awkward translations this can fall into different categories, and from our analysis, we classify them as *common sense, super sense, abstract phrase, idiom, metaphor, and ambiguity*, with ambiguity further sub-divided into context-unaware ambiguity, social/literature-unaware ambiguity, and coherence-unaware ambiguity. These classifications are to be further refined in the future, e.g., the differences between metaphor and idiom are sometimes fuzzy. We now list each of these with examples to support future MT research on improving the quality of MT when handling MWEs.

4.1.1 Common Sense

The first error category is called *common sense*. This category of error occurs when the translation models did not acquire any common knowledge or social context that human beings will take for granted, and that can be used to carry out reasoning wherever suitable. For instance, the sentence in Fig. 4 includes the vMWE “waved down” which in general understanding indicates that “he succeeded in getting the cab” and not only “waved his hand”. However, in the translation by DeepL and Bing this vMWE was wrongly translated as “he waved his hand to the cab” missing part of the original meaning; the MT

Source	Each time he took a walk, he felt as though he were leaving himself behind, and by giving himself up to the movement of the streets, by reducing himself to a seeing eye, he was able to escape the obligation to think, and this, more than anything else, brought him a measure of peace, a salutatory emptiness within.
DeepL	每一次散步，他都觉得自己好像把自己抛在了脑后，把自己交给了街道的运动，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这比什么都重要，这给他带来了一定程度的平静，内心的空虚让人敬佩。(jiāo gěi; nèi xīn de kōng xū ràng rén jìng pèi)
Bing	每次散步时，他都觉得自己好像要把自己抛在脑后，通过向街道运动自在，通过把自己留眼，他能够逃避思考的义务，这比什么都重要，给他带来了一定程度的和平，一种有益的空虚。(zì zài; yǒu yì de kōng xū)
Google	每次散步时，他都会感觉自己好像在抛弃自己，投身于大街小巷，睁大了眼睛，这使他摆脱了思考的义务，而这， 最重要的是，他带给他一种和平的感觉，一种内在的称呼空虚。(tóu shēn yú; nèi zài de chēng hu kōng xū)
Baidu	每次他散步，他都觉得自己好像是在离开自己，把自己交给街道的流动，把自己变成一个有眼光的人，他就可以逃避思考的义务，而这比其他任何事情都能给他带来某种程度的平和，一种内在的致意的空虚。(nèi zài de zhì yì de kōng xū)
Ref.	每一次散步，他都觉得自己好像把自己抛在了脑后，投身于大街小巷，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这给他带来了某种程度的平静和内心悦纳的空无，远胜于其他。(nèi xīn yuè nà de kōng wú)

Fig. 5 MT issues with MWEs: super sense

output by GoogleMT is also incorrect, saying "he waves with the cab in hand"; the Baidu translation of this sentence is semantically correct that "he waved and got one cab" though it does not use a corresponding Chinese side vMWE " 招手示停 (zhāo shǒu shì tíng)"[7].

4.1.2 Super Sense

This category of translation issue is related to *a form of state of mind* and we need to make a logical prediction to guess the correct interpretation, e.g. *positiveness*, *negativeness*, *neutral*, or other situations for some words, in the choice of Chinese characters. As in Fig. 5, the MT systems each have advantages for different parts of this long sentence. However, none of them is perfect. For instance, for the translation of vMWE "giving (himself) up (to)", the DeepL and Baidu outputs give very literal translation by saying "he gives himself to", the Bing translator drops the vMWE, while GoogleMT preserves the correct meaning in the translation "投身于 (tóu shēn yú)" from the reference indicating "he devoted himself". However, GoogleMT's output for the phrase "salutatory emptiness within" is very poor and makes no sense; the reference is "the emptiness that he welcomes" for which Baidu has a closer translation "内在的致意的空虚 (nèi zài de zhì yì de kōng xū)". All four MT outputs also use the same Chinese words "空虚 (kōng xū)" which is a term with a negative meaning, however, the sentence indicates that he is welcoming this emptiness, which should be the corresponding Chinese word "空无 (kōng wú)", an unbiased or positive meaning.

Source	Quinn had his doubts, but this was all he had to **go on**, his only bridge to the present.
DeepL	奎恩有他的疑虑，但这是他唯一的依靠，是他通往现在的唯一桥梁。(wéi yī de yī kào)
Bing	奎因有他的怀疑，但这就是他必须**去**的，他唯一的桥梁到现在。 (bì xū qù de)
Google	奎因有他的疑惑，但这就是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de)
Baidu	奎因有他的疑虑，但这是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de)
Ref.	奎恩曾有他的疑虑，但这是他**开展工作**的所有依据，是他通往现在的唯一桥梁。(**kāi zhǎn gōng zuò** de suǒ yǒu yī jù)

Fig. 6 MT issues with MWEs: abstract phrases

4.1.3 Abstract Phrases

The *abstract phrases* can have different exact meanings and we usually need some background information from the sentence or paragraph to select the correct word choices in the target language[8]. As mentioned in Section 3, in our post-editing and annotation task, the background context information of extracted sentences was given to the workers. With the example sentence in Fig. 6, from the context, we know that "go on" in this sentence means "to work from" using all the information he had. The phrase "this was all he had to go on" is then to be interpreted as "this is all the information he had to work from". At the end of the sentence, "the present" is the "present person" he needs to look for (with the picture of this person's younger age portrait). However, Bing translated it as "this is (where) he had to go" which is an incorrect interpretation of "had to go"; furthermore, Bing's translation of the second half of the sentence kept the English order, without any reordering between the words, which is grammatically incorrect in Chinese, i.e. "他唯一的桥梁到现在 (tā wéi yī de qiáo liáng dào xiàn zài)". GoogleMT and Baidu translated it as "what he need to do" which is also far from correct, while DeepL successfully translated the part "his only thing to relying on" but dropped the phrase "go on", i.e., *to do what*. *Abstract Phrase* can include *Super Sense* as its sub-category, however, it does not necessarily relate to a state of mind.

4.1.4 Idioms

The use of *idioms* often causes wrongly translated sentences, mostly resulting in humorous output due to literal translation. For example, in the sentence in Fig. 7, the vMWEs "*cutting capers*" and "tore back" are never translated correctly at the same time by any of the four MT models we used. The idiom

[7] We give full sentence pronunciation (in Pinyin) of Chinese characters in this figure, for the following examples, we only present the Chinese Pinyin for MWEs and studied words of the sentences to save space.

[8] it sometimes belongs to the context-unaware ambiguity (CUA) that we will mention later, however, CUA not necessarily means "abstract phrase", and usually needs paragraph information, not only sentence level. Furthermore, in some situations, we just don't know how to interpret "abstract phrase", i.e. the candidate interpretations are unknown without context, and this is different from ambiguity.

Source	I was smoking my pipe quietly by my dismantled steamer, and saw them all cutting capers in the light, with their arms lifted high, when the stout man with mustaches came tearing down to the river, a tin pail in his hand, assured me that everybody was 'behaving splendidly, splendidly, dipped about a quart of water and tore back again.
DeepL	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下高举双臂切着荠菜，这时，那个留着胡子的壮汉撕扯着来到河边，手里拿着一个铁皮桶，向我保证说，大家都'表现得很精彩，很精彩，浸了大约一夸脱的水，又撕扯回来了。 (qiē zhe jì cài; sī chě huí lái)
Bing	我悄悄地用我拆除的蒸笼抽我的烟斗，看到他们所有的切割披肩在光中，他们的手臂抬起来高，当粗壮的人用胡子来到河中，一个锡桶在他的手中，向我保证，每个人都'表现辉煌，辉煌，浸约夸脱的水，并再次撕毁。(qiē gē pī jiān; zài cì sī huǐ)
Google	我正用拆卸的蒸锅安静地抽着烟斗，看到他们都在灯光下割下刺山柑，双臂高高举起，当时那留着小胡子的矮壮男人向河边撕下，手里拿着一个铁桶，向我保证 每个人都表现得非常出色，表现得非常出色，蘸了一夸脱的水，然后又撕了回来。(gē xià cì shān gān; sī le huí lái)
Baidu	我正静静地在我拆掉的汽船旁抽着烟斗，看见他们都高举着胳膊，在灯光下割着山柑，这时那个长着胡子的胖男人手里拿着一个锡桶，朝河里跑来，向我保证每个人都表现得很好，很漂亮，蘸了一夸脱水，然后又往回跑。(gē zhe shān gān; wǎng huí pǎo)
Ref.	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下欢呼雀跃，高举双臂，这时，那个留着胡子的大块头，手里拿着一个铁皮桶，快速来到河边，向我确保大家都"表现得很精彩，很精彩"，他浸了大约一夸脱的水，又快速回去了。(huān hū què yuè; kuài sù huí qù)

Fig. 7 MT issues with MWEs: idioms

Source	The what? Auster laughed, and in that laugh everything was suddenly blown to bits. The chair was comfortable, and the beer had gone slightly to his head.
DeepL	那个什么？奥斯特笑了，在这笑声中，一切突然被炸得粉碎。(bèi zhà dé fěn suì) 椅子很舒服，啤酒已经微微到了他的头上。 (wēi wēi dào le tā de tóu shàng)
Bing	什么？奥斯特笑了，在笑，一切都突然被吹成位。(bèi chuī chéng wèi) 椅子很舒服，啤酒稍微到他的头去了。 (shāo wēi dào tā de tóu qù le)
Google	什么啊 Auster笑了起来，在那笑声中，一切突然被炸碎了。 (bèi zhà suì le) 椅子很舒服，啤酒微微飘到他的头上。(wēi wēi piāo dào tā de tóu shàng)
Baidu	什么？奥斯特笑了，在那笑声中，一切都突然被炸成碎片。(bèi zhà chéng suì piàn) 椅子很舒服，啤酒已经稍稍流到他的头上了。(shāo shāo liú dào tā de tóu shàng le)
Ref.	那个什么？奥斯特笑了，在这笑声中，一切突然化为乌有。(huà wéi wū yǒu) 椅子很舒服，啤酒已经微微让他上了头。 (wēi wēi ràng tā shàng le tóu)

Fig. 8 MT issues with MWEs: metaphor

"cutting capers" indicates frolic or romp, to "act in the manner of a young goat clumsily frolicking about" [9], and here it means "they are in a happy mood, playful and lively movement" which should properly be translated as the corresponding Chinese idiom "欢呼雀跃 (huān hū què yuè, happily jumping around like sparrows)". However, all four MT models translated it literally into "cutting" actions just with different objects, i.e., what they cut. The idiom (slang) "tore back" means the stout man *walked back rapidly*, which the Baidu translation gives the closest translation as "往回跑 (wǎng huí pǎo, run back)" but the other three models translated into an action "tear something (to be broken)" which is incorrect.

[9] https://www.dictionary.com/browse/cut-capers

4.1.5 Metaphors

The first sentence vMWE "blown to bits" in Fig. 8 is a *metaphor* to indicate "everything is gone", instead of the physical "blowing action". However, the three MT models DeepL, GoogleMT, and Baidu translate it as "exploded into pieces (by bombs)", while BingMT translates it even more literally into "blown to (computer) bits". There is a corresponding Chinese vMWE "化为乌有 (huà wéi wū yǒu, vanish into nothing)" which would be a proper choice for this source vMWE translation.

The second sentence vMWE "gone (slightly) to his head" is a metaphor to indicate "got slightly drunk". However, all four MT models translate it as physically "beer moved to his head" but by slightly different means such as *flow* or *flutter*. The corresponding translation as a MWE should be "微微让他上了头 (wéi wéi ràng tā **shàng le tóu**)", using the same characters, but the character order here makes so much difference, meaning "slightly drunk".

4.1.6 Ambiguity

We summarised different kinds of situations that cause ambiguity in the resulting translation when it meets MWEs or named entities, so we further divide ambiguity into three sub-classes.

I).Context-Unaware Ambiguity

In this case, the *context*, i.e. the background information, is needed for the correct translation of the sentence. For instance, see Fig. 9. DeepL gives the translation "it did not give me time though", while Bing and GoogleMT give the same translation "it/this did not give me one day's time" and Baidu outputs a grammatically incorrect sentence. From the pre-context, we understand that it means the speaker "did not feel that is special to him" or "did not have an affection of that" after *all the Mormon missionary's effort towards him*. Interestingly, there is a popular Chinese idiom (slang) that matches this meaning very well "不是我的菜 (bù shì wǒ de cài, literally *not my dish*)". We also offer an alternative translation in our corpus for this sentence as "但是他没有关注我 (dàn shì tā méi yǒu guān zhù wǒ, but he did not pay attention to me)" since there is this idiom "not to give someone the time of day" indicating "not to pay attention to someone" and it also makes sense in this story context. From this point of view, the context-based MT model deserves some more attention, instead of only focusing on sentence level. When we tried to put all background context information as shown in Fig. 9 into the four MT models, they produce the same output for this studied sentence, as for sentence level MT. This indicates that current MT models still focus on sentence-by-sentence translation when meeting paragraphs, instead of using context inference; or alternatively, that even if the system is able to utilise multi-sentence information, that information is not sufficient to resolve the correct translation at least in this case.

II).Social/Literature-Unaware Ambiguity:

In this case, *social knowledge* of current affairs from news, or *literature knowl-*

Source	But it did not give me the time of day.
DeepL	但它并没有给我时间。 (gěi wǒ shí jiān)
Bing	但它没有给我一天的时间。 (gěi wǒ yī tiān de shí jiān)
Google	但这没有给我一天的时间。 (gěi wǒ yī tiān de shí jiān)
Baidu	但它没有给我一天中的时间。 (gěi wǒ yī tiān zhōng de shí jiān)
Ref.	但我没有感到这个对于我特殊 / 但这不是我的菜。(gǎn dào zhè ge duì yú wǒ tè shū / … wǒ de cài) 但是他没有关注我 (dàn shì tā méi yǒu guān zhù wǒ, but he did not pay attention to me)
Context	An old Mormon missionary in Nauvoo once gripped my knee hard as we sat side by side, and he put his arm about me and called me "Brother." We'd only met ten minutes before. He took me to his good bosom. His eyes began to mist. I was a prospect, an exotic prospect in old tennis shoes and a sweatshirt. His heart opened to me. It opened like a cuckoo clock. But it did not …

Fig. 9 MT issues with MWEs: context-unaware ambiguity

Source	The moment they know the de-gnoming's going on they storm up to have a look. Then someone says that it can't be long now before the Russians write Arafat off.
DeepL	他们一知道去核的事，就会冲上去看一看。 (qù hé) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**注销**。(**bǎ** ā lā fǎ tè **zhù xiāo**)
Bing	当他们知道去诺格明是怎么回事， 他们冲了起来看看。 (qù nuò gé míng) 然后有人说，现在俄罗斯人要不长了，就**把**阿拉法特**注销**了。(**bǎ** ā lā fǎ tè **zhù xiāo** le)
Google	当他们知道正在逐渐消失的那一刻，他们便冲上去看看。 (zhèng zài zhú jiàn xiāo shī) 然后有人说，不久之后俄罗斯人**将**阿拉法特**注销**。 (**jiāng** ā lā fǎ tè **zhù xiāo**)
Baidu	他们一知道德格诺明正在进行，就冲上去看一看。 (dé gé nuò míng) 然后有人说，俄国人很快就会把阿拉法特**一笔勾销**了。 (bǎ ā lā fǎ tè **yī bǐ gōu xiāo** le)
Ref.	一知道去地精的事在进行，他们就冲上去观看。 (qù dì jīng) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**下課 /** 让…下台。 (**bǎ** ā lā fǎ tè **xià kè;** ràng…xià tái)

Fig. 10 MT issues with MWEs: social/literature-unaware ambiguity

edge about some newly invented entities and phrases are required in order to get a correct translation output. For instance, Fig. 10 includes two sentences, one from politics and another from literature.

In the first sentence, "de-gnoming" is a literature word from Harry Potter, invented by its author, to refer to the process of ridding a garden of gnomes, *a small magical beast*. Without this literature knowledge, it is not possible to translate the sentence correctly. For instance, even though this sentence is from a very popular novel that has been translated into most languages, DeepL translated it as "去核 (qù hé, de-nuclear)", Bing translated it as "去诺格明 (qù nuò gé míng, *de*-nuògémíng" where "nuògémíng" is a simulation of the pronunciation of "gnoming" in a Chinese way, Baidu translated it as "德格诺明 (dé gé nuò míng)" which is the simulation of the pronunciation of the overall term "de-gnoming". As we mentioned in an earlier section on corpus preparation, we add examples of translation pairs where we think it is a challenge to MT, including words with corresponding translations as MWEs in the target language. In this sentence, we treat the target Chinese translation "去地精 (qù dì jīng)"as an MWE with a special structural construction "去 (qù, *get rid of*) + noun"indicating "remove/delete + noun". Since the word

Source	Two months ago I had to have an operation for a serious ***complaint***.
DeepL	两个月前，我因为一次严重的**投诉**不得不做手术。(**tóu sù** ... zuò shǒu shù)
Bing	两个月前，我不得不做一个严重的**投诉**手术。 (zuò ... **tóu sù** shǒu shù)
Google	两个月前，我不得不接受一次手术以应对严重的**投诉**。 (jiē shòu yī cì shǒu shù ... **tóu sù**)
Baidu	两个月前，我因为严重的**投诉**不得不动手术。 (**tóu sù** ... dòng shǒu shù)
Ref.	两个月前，我因为一次严重的**症状**不得不做手术。(**zhèng zhuàng** ... zuò shǒu shù)

Fig. 11 MT issues with MWEs: coherence-unaware ambiguity

"de-gnoming"is a literature word, the Chinese MWE "去地精 (qù dì jīng)"is also regarded as a borrowed foreign word. The original vMWEs annotated in this English sentence are "going on"(VPC.full) and "have (a) look"(LVC.full); however, in the MT task, apparently these two vMWEs have been addressed better by MT engines than the literature word "de-gnoming", even though the Bing translator also made a mistake on translating "going on" into " 是怎么回事 (shì zěn me huí shì, what is going on)".

In the second sentence, "write Arafat off" is to dismiss "Yasser Arafat", Chairman of the Palestine Liberation Organization, which is a historical person's name. However, all three models DeepL, Bing, and GoogleMT translated it into "把/将阿拉法特注销 (bǎ/jiāng ā lā fǎ tè zhù xiāo, *deregister Arafat*)" which treated "Arafat" as a title of certain policy/proceeding, not being able to recognize it as a personal named entity, while Baidu made the effort to use the Chinese idiom "一笔勾销 (yī bǐ gōu xiāo, *cancel everything*, or *never mention historical conflicts*)" for "write off" but it is not a correct translation. Interestingly, if we put these two sentences into a web search engine it retrieves the correct web pages as the context in the top list of the search result. This may indicate that future MT models could consider including web search results as part of their knowledge of background for translation purposes, e.g. using generative translation models [24, 25].

III).Coherence-unaware ambiguity

This kind of MWE ambiguity can be solved by the *coherence* of the sentence itself, for instance, the example in Fig. 11. The four MT models all translated the vMWE itself "have an operation" correctly in meaning preservation by "做/接受/动手术 (zuò/jiē shòu/dòng shǒu shù)" just with different Chinese word choices. However, none of the MT models translated the "reason of the operation", i.e., "complaint" correctly. The word complaint has two most commonly used meanings "a statement that something is unsatisfactory or unacceptable" or "an illness or medical condition" and all four models chose the first one. According to the simple logic of social life, people do not need to "have an operation" due to "a statement", instead their "medical condition" should have been chosen to translate the word "complaint". Because of the incorrectly chosen candidate translation of the word "complaint", Bing's output even invented a new term in Chinese "投诉手术 (tóu sù shǒu shù, *a surgery of*

complaint statement kind)" which makes no sense. This raises the issue of the "cognitive plausibility" of current powerful language models in NLP fields.

4.2 *English-to-German*

In the case of English-to-German MWEs, there are some cases where the corresponding German translation of English MWEs can be one word. This is partial because German has compound verbs. For instance, the vMWE "woke up" the sentence "An old woman with crinkly grey hair woke up at her post outside the lavatory and opened the door, smiling and grasping a filthy cleaning rag." has corresponding German aligned word "erwachte" with a suitable translation "Eine alte Frau mit krausem, grauem Haar erwachte auf ihrem Posten vor der Toilette und öffnete die Tür, lächelte und griff nach einem schmutzigen Putzlappen.".

This "MWE to single-word" alignment phenomenon might not be specific for the English-German language pair, since this also occurs in the English-to-Chinese and English-to-Arabic translation, such as an English verb+particle MWE being aligned to one single Chinese character/word. For example, in this sentence "The fact that my name has been mixed up in this.", the vMWE (VPC) *mixed up* gets aligned to a single character word "混 (hùn)" in a suitable translation "事实上，我的名字已经被混在这里面了。(shì shí shàng, wǒ de míng zì yǐ jīng bèi hùn zài zhè lǐ miàn le)". However, in the view that German has many compound words and it might be more frequent than other languages, we list this issue under the English-German sub-category.

A second issue is that the automatic translation to German can be very *biased* towards choosing the formal (or polite) form *vs* informal form. See the examples such as "Sie"instead of the second form singular "du"for "you", "auf Basis von" instead of "basierend auf" for "based on". To achieve a higher level of translation accuracy, MT models need to choose the correct or more suitable form of words depending on the context being used.

A third issue is that, for the English verbal multi-word expressions that are often not translated as verbal multi-word expressions to German, this indicates some further work to explore by MT researchers to develop better models to have the machine producing corresponding German existing MWEs.

4.3 *English-to-Polish*

The MT output issues in English-to-Polish mostly fall into the categories of coherence-unaware error, literal translation, context-unaware error, and gender-related mistakes, in addition to other errors similar to what we listed in the Chinese-English category.

Regarding the MT output issues on English-to-Polish that fall into coherence-unaware error, for instance, the vMWE "write off" in the sentence "Then someone says that it can't be long now before the Russians write Arafat off." was

translated as “Wypiszą” (Potem ktoś mówi, że już niedługo Rosjanie wypiszą Arafata.) which means “prescribe”, instead of the correct “spiszą na straty (Arafata)”. This error can be avoided by the coherence of the sentence itself in meaning preservation models.

For the literal translation, we can see the example vMWE “gave (him) a look” in the sentence “She ruffled her feathers and gave him a look of deep disgust.” which was literally translated as “dała mu spojrzenie”, however, in Polish, people use “**throw** a look” as “rzuciła (mu) spojrzenie” instead of “gave (dała, a female form)”[10]. Another example of literal translation leading to errors is the vMWE “turn the tables” from the sentence “Now Iran wants to turn the tables and is inviting cartoonists to do their best by depicting the Holocaust.” which is translated as “odwrócić stoliki (turn tables)”, however, it should be “odwrócić sytuację (turn the situation)” or “odwrócić rolę (turn role)” with a proper translation “*Teraz Iran chce odwrócić sytuację i zachęca rysowników, by zrobili wszystko, co w ich mocy, przedstawiając Holocaust.*” These two examples illustrate the localisation issue in the target language.

For the context unaware issue, we can look back to the example sentence “But it did not give me the time of day.” from Fig. 9. This was literally translated word by word into “Ale nie dało mi to pory dnia.” which is in the sense of hour/time. However, it should be “Nie sądzę aby to było coś wyjątkowo/szczególnie dla mnie. (I do not think this is special to me.)” based on the context, or “Ale to nie moja bajka” as an idiomatic expression which means “not my fairy tale” (indicating *not my cup of tea*).

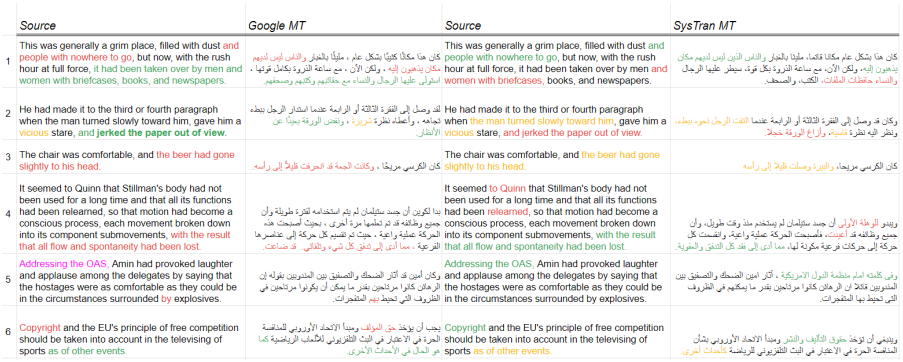
Regarding gender-related translation errors, for instance, in some translations, the word “I” in English shall be reconsidered if it specifies a male or female which results in different translations in the target Polish.

We listed some examples of mistranslation of gender-related terms in Fig. 25 (Section Appendix). The word “friend”in English can be both male and female, but in Polish male is “przyjaciel”and the female “przyjaciółka”. The example sentence “I have a friend”was translated using the male form “mam przyjaciela”instead of the female form “Mam przyjaciółkę”even though there is the context of “her”indicating the friend is a female gender. Also, the English word “that”in this situation is gender specific in Polish as well, because it leads to a clause relating to “friend”, and “przyjaciel”and “przyjaciółka”are conjugated (because of the verb “have”followed).

5 Quantifying MT Errors using English-Arabic for Case Study

For Arabic corpus creation, firstly there is no Arabic language option in the DeepL Translator. Following the same procedure on the corpus creation of other languages, by comparing MT systems and selecting one as raw output provider for next step post-editing, our two native Arabic speakers per-

[10] a proper translation: *Nastroszyła sobie pióra i rzuciła mu spojrzenie głębokiego obrzydzenia.* Also the MT output word for “Nastroszyła” was “Zdruzgotała” which has the wrong meaning.



	Source	Google MT	Source	SysTran MT
1	This was generally a grim place, filled with dust and people with nowhere to go, but now, with the rush hour at full force, it had been taken over by men and women with briefcases, books, and newspapers.	كان هذا مكانًا كئيبًا بشكل عام ، مليئًا بالغبار والناس الذين ليس لديهم مكان يذهبون إليه ، ولكن الآن ، مع ساعة الذروة بكامل قوتها ، استولى عليها الرجال والنساء مع حقائبهم وكتبهم وصحفهم.	This was generally a grim place, filled with dust and people with nowhere to go, but now, with the rush hour at full force, it had been taken over by men and women with briefcases, books, and newspapers.	كان هذا بشكل عام مكانا قاتما، مليئا بالغبار والناس الذين ليس لديهم مكان يذهبون إليه، ولكن الآن، مع ساعة الذروة بكل قوة، سيطر عليها الرجال والنساء بحافظات الملفات، الكتب، والصحف.
2	He had made it to the third or fourth paragraph when the man turned slowly toward him, gave him a vicious stare, and jerked the paper out of view.	لقد وصل إلى الفقرة الثالثة أو الرابعة عندما استدار الرجل ببطء تجاهه ، وأعطاه نظرة شريرة ، وانتزع الورقة بعيدًا عن الأنظار.	He had made it to the third or fourth paragraph when the man turned slowly toward him, gave him a vicious stare, and jerked the paper out of view.	وكان قد وصل إلى الفقرة الثالثة أو الرابعة عندما التفت الرجل نحوه ببطء، ونظر اليه نظرة شنيعة، وأزاع الورقة خجلا.
3	The chair was comfortable, and the beer had gone slightly to his head.	كان الكرسي مريحًا ، وكانت الجعة قد انحرفت قليلاً إلى رأسه.	The chair was comfortable, and the beer had gone slightly to his head.	كان الكرسي مريحا، والبيرة وصلت قليلا إلى رأسه
4	It seemed to Quinn that Stillman's body had not been used for a long time and that all its functions had been relearned, so that motion had become a conscious process, each movement broken down into its component submovements, with the result that all flow and spontaneity had been lost.	بدا لكوين أن جسد ستيلمان لم يتم استخدامه لفترة طويلة وأن جميع وظائفه قد تم تعلمها مرة أخرى ، بحيث أصبحت هذه الحركة عملية واعية ، حيث تم تقسيم كل حركة إلى عناصرها الفرعية ، مما أدى إلى تدفق كل شيء والتلقائي. قد ضاعت.	It seemed to Quinn that Stillman's body had not been used for a long time and that all its functions had been relearned, so that motion had become a conscious process, each movement broken down into its component submovements, with the result that all flow and spontaneity had been lost.	ويبدو للرجلة الأولى أن جسد ستيلمان لم يستخدم منذ وقت طويل، وأن جميع وظائفه قد أعيدت، فأصبحت الحركة عملية واعية، وانقسمت كل حركة إلى حركات فرعية مكونة لها، مما أدى إلى فقد كل التدفق والعفوية.
5	Addressing the OAS, Amin had provoked laughter and applause among the delegates by saying that the hostages were as comfortable as they could be in the circumstances surrounded by explosives.	وكان أمين قد أثار الضحك والتصفيق بين المندوبين بقوله إن الرهان كانوا مرتاحين بقدر ما يمكن أن يكونوا مرتاحين في الظروف التي تحيط بهم المتفجرات.	Addressing the OAS, Amin had provoked laughter and applause among the delegates by saying that the hostages were as comfortable as they could be in the circumstances surrounded by explosives.	وفي كلمته امام منظمة الدول الامريكية ، أثار امين الضحك والتصفيق بين المندوبين قائلا ان الرهائن كانوا مرتاحين بقدر ما يمكنهم في الظروف التي تحيط بها المتفجرات.
6	Copyright and the EU's principle of free competition should be taken into account in the televising of sports as of other events.	يجب أن يؤخذ حق المؤلف ومبدأ الاتحاد الأوروبي للمنافسة الحرة في الاعتبار في البث التلفزيوني للألعاب الرياضية كما هو الحال في الأحداث الأخرى.	Copyright and the EU's principle of free competition should be taken into account in the televising of sports as of other events.	وينبغي أن تؤخذ حقوق التأليف والنشر ومبدأ الاتحاد الأوروبي بشأن المنافسة الحرة في الاعتبار في البث التلفزيوني للرياضة كالمسابقات الجرى.

Fig. 12 MT Output Comparisons between GoogleMT and Systran (Green: well translated, Red: wrong translation, Yellow: correct but unnatural and Magenta: skipped.)

formed some manual comparison on GoogleMT versus SysTran MT [11] and we show some working examples in Figure 12. In these examples, we find that GoogleMT produced more *meaning*-correct outputs in comparison to the SysTran engine, even though GoogleMT made more mistakes on *entity* translations. We thought entity errors are easier to correct and this procedure also gives us the chance to look into MWE-related errors by MT engines. So we picked GoogleMT for Arabic corpus creation.

5.1 *Revisiting HOPE Metric*

To qualitatively and quantitatively evaluate the MT output of English-to-Arabic translation, i.e. how many errors the state-of-the-art MT system makes in this language pair on our data and to what severity levels, we deployed the HOPE metric [6] for this purpose. There are eight original error types designed in HOPE, which include Mistranslation (MIS), Style (STL), Terminology (TRM), Impact (IMP), Missing Required Adaptation, Ungrammatical (UGR), Proofreading Error (PRF), and Proper Name. To reflect our task more precisely, we added two more error types during our post-editing and scoring, which are "MWE Missed Chance (**MMC**)" and "Skipped Word (**SKP**)". The MMC error is to quantify the situation when the source-side MWEs are translated either wrongly in the target, or correctly in meaning but the model missed the chance of choosing the correct target MWEs when there are indeed such MWEs in the standard Arabic language. The SKP error is to highlight the situation when GoogleMT skipped some important source words during the translation procedure.

The HOPE metric has a hierarchical scoring procedure, from the segment level to the system level. Each error type is also assigned with specific penalty scores according to the severity level, using 2^n (1, 2, 4, 8, 16) for "minor,

[11] `https://www.systransoft.com/lp/machine-translation/`

Types	MMC	MIS	STL	TRM	IMP	UGR	PRF	SKP	All	PPS
T.P.S.	76	68	69	39	114	37	46	6	455	
R.o.T.	17%	15%	15%	9%	25%	8%	10%	1%		3.03

Table 1 Detailed Error Types and Ratios (T.P.S.: total penalty scores. R.o.T.: Ratio out of total segments.)

medium, major, severe, and critical" levels. The segment-level score is the sum of each individual error score. To classify the errors in a more straightforward view, the segment-level scores are grouped into three categories, either minor (score 1-4), major (score 5+), or correct.

There are a few key indicators in HOPE metric. 1) The total penalty score (T.P.S.) for each type of error is the sum of all penalty scores overall segments. 2) The ratio of total segments (R.o.T.) is the percentage of specific error types out of all errors T.P.S. 3) The PPS value means the penalty point per segment that is the value of all penalty scores T.P.S. divided by segment numbers, i.e. $PPS = Sum_{penalty}/Sum_{segment}$.

5.2 *HOPE Scoring Output*

The HOPE error score marking was performed simultaneously when the post-editing was carried out for each segment. To reflect details on different error types, we list the total penalty scores (T.P.S.) and their ratio (R.o.T.) out of all segments in Table 1 using 150 segments. The T.P.S. value is 455 and the number of segment is 150. From this detailed error type analysis, we can see that the MMC and SKP error types that we added to the HOPE metric occupied 17 percent and 6 percent of errors. This also reflected that MWE-related errors take nearly 20 percent of all errors in this testing output. This finding implicitly implies the importance of our multilingual corpus creation with MWE annotations that can be used for MT model validations.

The statistics of GoogleMT output on error categories are shown in Figure 13, which tells that 21 percent of the output segments fall into major errors that need to be fixed, 44 percent of segments fall into minor errors that can be good enough for some applications, and only 35 percent of them are correct translations that do not need to edit.

6 Discussion

In this section, we present some broader issues that we discovered during our corpus construction process, which are related to MWEs and MT.

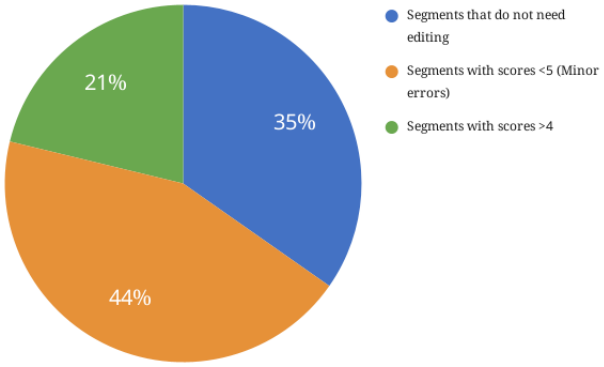


Fig. 13 Quantitative Evaluation on GoogleMT Outputs from English to Standard Arabic using HOPE Metric: Minor Errors *vs* Major Errors

6.1 *Issues in the source English corpus*

Some problems occurred in the source English corpus which impact the sentences we extracted from the monolingual training and test data with vMWE annotation tags.

Firstly, there is an accuracy issue with the tagging, which may have been introduced as a result of the tagging task being carried out with one person per sentence. Each sentence, either long or short, is located in one line. [12].

Some error annotations of vMWEs in the source monolingual corpus will thus have some impact on the accuracy level of the *vMWE discovery and identification* shared task but also affect the bilingual usage of AlphaMWE, so we tried to address all these cases. For instance, in the example sentence in Fig. 6, the English corpus annotated wrongly the sequence "had to go on" as a verbal idiom (VIDs) which is not accurate. The verb "had" here is affiliated with "all he had" instead of "to go on". So either we annotate "go on" as vMWE in the sentence or the overall clause "*all he had to go on*" as a studied term. In AlphaMWE, we reserved "(to) go on"as the vMWE studying term, since "all he had to go on"is more like a sentence clause than a vMWE term.

Another example with a different type of vMWE is the sentence "He put them on in a kind of trance." where the source English corpus tagged "put" and "trance" as Light-verb construction (LVC.cause). However, the phrase is with "put...on" instead of "put...trance". The phrase "put someone into a trance" is to express "make someone into a half-conscious state". However, for this sentence, if we check back a bit further of the context, it means "he put on his cloth in a kind of trance". The word "trance" is affiliated with the phrase "*in a kind of trance*" instead of "put".

A second issue we discovered with the English language corpus is that there are some interesting sentences in the corpus that include non-decomposable

[12] https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.1/EN

Source	The pressure had been building up in him since Stillman's disappearance that morning, and *it came out of him now as* ***a torrent of words***.
DeepL	自从那天早上斯蒂尔曼失踪后，他身上的压力就一直在积聚，现在却如**滔滔江水**般涌了出来。(jī jù, … què rú **tāo tāo jiāng shuǐ** bān yǒng le chū lái)
Bing	自从斯莱特曼那天早上失踪后，他的压力就一直在增加，现在他突然**大言不出**。(zēng jiā, … tā tú rán **dà yán bù chū**)
Google	自从斯蒂尔曼（Stillman）那天早上失踪以来，他身上的压力一直在增加，现在，他的**言语如潮水**般涌出。(zēng jiā, … tā de **yán yǔ rú cháo shuǐ** bān yǒng chū)
Baidu	自从斯蒂尔曼那天早上失踪后，他的压力就一直在增加，现在他身上的压力就像**滔滔不绝的话**。(zēng jiā, … tā shēn shang de yā lì jiù xiàng **tāo tāo bù jué de huà**)
Ref.	自从那天早上斯蒂尔曼失踪后，他身上的压力就一直在积聚，现在这些压力变成**滔滔不绝的话**倾诉而出。(jī jù, … zhè xiē yā lì biàn chéng **tāo tāo bù jué de huà** qīng sù ér chū)

Fig. 14 Additional vMWEs or MWEs

MWEs but these MWEs are not annotated. We plan to add further annotation on this aspect and extend this kind of bilingual pairs.

For instance, in Fig. 14, the vMWE category verb-particle constructions (VPC.semi) is tagged to the phrase "building up", which may be an interesting case for vMWE discovery and identification, however, for cross-lingual research, as the initial aim of our corpus construction, such as MT, the ending part of this sentence "came out of him now as a torrent of words" poses more challenges, and would draw more attention from researchers.

We test this with four MT models and found the following outputs: DeepL literally translated it into "a strong and fast-moving stream of water"and dropped "words"; Bing gave a translation with the opposite meaning "does not say a word"; Google and Baidu produced much better translation covering "torrent of words"even though the sentence level translation contain errors and can be improved, i.e., "it"meaning the "pressure"in the source sentence was dropped out by Google; "came out of him"was dropped out by Baidu.

6.2 *Broader findings on MT issues with MWEs*

In this section, we introduce some broader findings on MT issues and interesting phenomena related to MWEs that still exist for the current state-of-the-art models, focusing on the English-to-Chinese pair.

6.2.1 Entity Translation Issue

Named entities (NEs) from Western languages always match into multi-character / word expressions in the Chinese language, even though it is a single word in the original Western form, e.g. the family name or surname of a person. In some documentary stories with some famous names, including ones from Bible, the mistranslation of named entities will lead to inaccurate history, and possibly cause arguments. For instance, from the Bible, *Absalom* was the son of David, the King of Israel, while *Abraham* was the founding father of the

Source	The Herodian relics are all that relics should be columns distorted, well worked over by time, Absalom's tomb with its bulbous roof and odd funnel tapering out of it.
DeepL	希律王的遗迹是所有的遗迹都应该是柱子扭曲的，被时间加工过的好的，亚伯拉罕的墓，它的屋顶是球状的，奇怪的漏斗从里面渐渐露出来。(xī lǜ wáng …,… yà bó lā hǎn)
Bing	英雄遗物都是文物，应该被柱子扭曲，随着时间的推移，Absalom的坟墓与它的球状屋顶和奇怪的漏斗逐渐减少出来。(yīng xióng…, … Absalom)
Ref.	希律王的遗迹都是柱子成扭曲状的，随着时间的推移精心加工，押沙龙式的坟墓带有球茎状的屋顶和奇怪的漏斗渐渐地从里面露出来。(xī lǜ wáng …,… yā shā lóng)

Fig. 15 General MT: named entity

"Covenant of the pieces", the special relationship between the Hebrews and God [13]. In this example sentence in Fig. 15, the named entity "Herodian"was correctly translated by DeepL as "希律王 (xī lǜ wáng)", however, wrongly translated as "hero"by Bing only taking part of this word "Hero" instead of "Herod" from Herod the Great. The named entity "Absalom"was kept as an unknown word by Bing without translation, and was wrongly translated by DeepL into "亚伯拉罕 (yà bó lā hǎn)", which is actually from another named entity "Abraham", and the correct translation of it in Chinese is the multi-character/word sequence "押沙龙 (yā shā lóng)"[14].

6.2.2 The English-Style Chinese Patterns

The English-style Chinese situation occurred when the MT models face some translation of English phrase patterns. We name this as *English-style Chinese* or *EngliChinese* in short, which means the Chinese sentence sounds like the English pattern-based expression, such as either the syntax of the Chinese sentence is in English, or there are some English input words (or literally translated words) that apparently make the sentence sound awkward in Chinese.

For the first example sentence, Fig. 16 "will find ourselves (forced to) ⋯ "is an English pattern that shall not be translated just word by word into Chinese. However, DeepL and Google both made the same mistake by the literal translation into Chinese, either dropping "find ourselves"in the translation or replacing "find"as "realize"which would make more sense in Chinese, such as "将 (意识到自己) 不得不, jiāng (yì shí dào zì jǐ) bù dé bù". The second sentence, "see the last of its Christianity"contains one idiom "see the last of (someone)", however, DeepL, Bing, and Baidu all translated it into an awkward Chinese sequence, especially Baidu, which literally translated it word by word from English plus a literal moving of "its"to the front.

[13] https://www.thebiblejourney.org
[14] https://biblehub.com http://biblehub.net/searchchin.php?q= 押沙龙

Source	Otherwise we will find ourselves forced to take more serious decisions as regards our trading relations with Israel. At times I suspect that the world would be glad to **see the last of its Christianity**, and that it is the persistency of the Jews that prevents it.
DeepL	否则，我们将发现我们将被迫在与以色列的贸易关系上作出更严肃的决定。(jiāng fā xiàn wǒ men jiāng bèi pò) 有时我怀疑，世界会很高兴**看到其最后的基督教**，而阻止它的是犹太人的顽固不化。(kàn dào qí zuì hòu de jī dū jiào)
Bing	否则，我们将被迫就我们与以色列的贸易关系作出更严肃的决定。(jiāng bèi pò jiù) 有时我怀疑，世界会很高兴**看到其基督教的最后一个**，是犹太人的坚持阻止它。(kàn dào qí jī dū jiào de zuì hòu yī gè)
Google	否则，我们将发现自己被迫就与以色列的贸易关系做出更严肃的决定。(jiāng fā xiàn zì jǐ bèi pò) 有时我怀疑，世界将会很高兴**看到基督教的最后一幕**，而犹太人的顽强阻止了它的发展。(kàn dào jī dū jiào de zuì hòu yī mù)
Baidu	否则，我们将被迫对我们与以色列的贸易关系作出更严肃的决定。(jiāng bèi pò duì) 有时我怀疑，世界会很高兴**看到它的基督教的最后一次**，是犹太人的坚持阻止了它。(kàn dào tā de jī dū jiào de zuì hòu yī cì)
Ref.	否则，我们将(意识到自己)不得不就与以色列的贸易关系作出更严肃的决定。(jiāng *yì shí dào zì jǐ* bù dé bù jiù) 有时我怀疑，世界会很高兴**跟基督教说拜拜**，而阻止它的是犹太人的坚持。(gēn jī dū jiào shuō bài bài)

Fig. 16 General MT: *English-style* Chinese

Source	We step into the street and my friend David Shahar, whose chest is large, takes a deep breath and advises me to do the same.
DeepL	我们步入街头，我的朋友大卫-沙哈尔**胸大无脑**，他深吸一口气，建议我也这样做。(xiōng dà wú nǎo)
Google	我们走进这条街，我的朋友大卫·沙哈（David Shahar）的**胸大了**，深吸了一口气，建议我也这样做。(xiōng dà le)
Ref.	我们步入街头，我的朋友大卫-沙哈胸膛大，他深吸一口气，建议我也这样做。(xiōng táng dà)

Fig. 17 General MT: misusage of MWEs

6.2.3 Mis-use of MWEs

This issue is also related to MWEs. The current SOTA MT models are apparently able to incorporate some target side MWEs into their candidate output. However, some MWEs on the target side were wrongly picked up by the MT engine which resulted in the translation having a different meaning to the source.

The output of DeepL with the example sentence in Fig. 17 used a Chinese idiom "胸大无脑 (xiōng dà wú nǎo, literally "big chest, no brain") to translate "whose chest is large"which is wrong by virtue of adding some extra meaning that does not exist in the source text. This may be caused by the training corpora that DeepL used which have a higher probability to align "chest ... large ..."to "胸大无脑(xiōng dà wú nǎo)". In addition, the Google output is not correct either by using "胸大了 (xiōng dà le, chest becomes large)"which is different from the original meaning.

Source	Still, it was painful, and he struggled desperately to swallow his pride.
DeepL	sentMT: 但还是很痛苦，他拼命挣扎着吞下自己的骄傲。(tūn xià zì jǐ de jiāo'ào) docMT: 不过，这还是很痛苦的，他拼命挣扎着吞下自己的自尊心。(tūn xià zi jǐ de zì zūn xīn)
Google	sentMT: 仍然，那是痛苦的，他拼命挣扎着吞下自己的骄傲。(tūn xià zì jǐ de jiāo'ào) docMT: 尽管如此，这还是很痛苦的，他拼命挣扎着吞下自己的骄傲。(tūn xià zì jǐ de jiāo'ào)
Ref.	尽管这很痛苦，他拼命挣扎着忍下自己的自尊心（没有做回应）。(rěn xià zì jǐ de zì zūn xīn *méi yǒu zuò huí yīng*)
Context	He was about to tell her who he was, but then he realized that it made no difference. The girl was beyond hope. For five years he had kept William Wilson's identity a secret, and he wasn't about to give it away now, least of all to an imbecile stranger. *Still, it was painful, and he struggled desperately to swallow his pride.* Rather than punch the girl in the face, he abruptly stood up from his seat and walked away.
Source	Shrugging, he gives up and I turn to the twice disagreeable chicken and eat guiltily, my appetite spoiled.
DeepL	sentMT: 耸了耸肩，他放弃了，我转身对着那只两次不欢而散的鸡，内疚地吃着，胃口被破坏了。(nà zhǐ liǎng cì bù huān ér sàn de jī) docMT: **耸耸肩，他放弃了，我转而吃起了两次不喜欢的鸡肉，内疚地吃着，胃口被破坏了。**(liǎng cì bù xǐ huān de jī ròu)
Google	sentMT: 耸了耸肩，他放弃了，我转向那两次令人讨厌的鸡肉，内eat地吃了，我的胃口变坏了。(nà liǎng cì lìng rén tǎo yàn de jī ròu) docMT removing # text =: 耸耸肩，他放弃了，我转向那两次令人讨厌的鸡肉，内地吃了，我的胃口变坏了。(nà liǎng cì lìng rén tǎo yàn de jī ròu)
Ref.	他耸了耸肩，放弃了，我转身对着那只引起两次争论的鸡，内疚地吃着，胃口变坏了。(nà zhǐ yǐn qǐ liǎng cì zhēng lùn de jī)
Context	You must never never eat trephena food again." "I can't promise you that. You're asking too much. And just for one sandwich." "I have a duty toward you," he tells me. "Will you listen to a proposition?" "Of course I will." "So let us make a deal. I am prepared to pay you. If you will eat nothing but kosher food, for the rest of your life I will send you fifteen dollars a week." "That's very generous," I say. "Well, you are a Jew," he says. "I must try to save you." "How do you earn your living?" "In a Hasidic sweater factory in New Jersey. We are all Hasidim there. The boss is a Hasid. I came from Israel five years ago to be married in New Jersey. My rabbi is in Jerusalem." "How is it that you don't know English?" "What do I need English for? So, I am asking, will you take my fifteen dollars?" "Kosher food is far more expensive than other kinds," I say. "Fifteen dollars isn't nearly enough." "I can go as far as twenty-five." "I can't accept such a sacrifice from you." *Shrugging, he gives up and I turn to the twice disagreeable chicken and eat guiltily, my appetite spoiled*.

Fig. 18 General MT: document-level MT (docMT) vs sentence-level MT (sentMT)

6.3 *Document-level MT for MWEs*

In the context-unaware ambiguity section, we mentioned document level MT as having some issues and here we give further analysis on that topic. There are some phrases and MWEs that usually need context information to understand correctly first, before translation. In some cases, it is ambiguity, while in others it is just out of the blue that we do not have a clue how to translate the terms or phrases being used, i.e. they are totally unexpected. We present two examples here, one vMWE as an idiom and another as a noun phrase (adj+adj+noun), in Fig. 18. The examples indicate that the current SOTA MT models can't handle this well yet.

For the first sentence, the idiom "swallow someone's pride"means "to decide to do something although it will make him/her feel embarrassed or ashamed", and in this context, it means he "did not say anything", instead, just walked away [15]. Here we can translate it as "忍下自己的自尊心 (rěn xià zìjǐ de zì zūn xīn)"or "没有做回应 (méi yǒu zuò huí yīng)". DeepL document level MT gave a good pick up using the word "自尊心 (zì zūn xīn)", however,

[15] https://dictionary.cambridge.org/dictionary/english/swallow-your-pride

"吞下自己的自尊心 (tūn xià zìjǐ de zì zūn xīn)"is not a correct way to say this in Chinese. All other three MT models, including sentence level DeepL, and sentence/document level GoogleMT yielded the same literal translation "吞下自己的骄傲 (tūn xià zìjǐ de jiāo ào)"which makes no sense in Chinese [16].

For the second sentence, from the context, we understand the point that "the twice disagreeable chicken"means that "the chicken caused twice disagreements". One happened in the beginning "You must never...I can't promise you that" and another happened in the end "So, I am asking, will you take my ...I can't accept such a sacrifice from you." However, the document-level MT of DeepL and Google both failed to translate this meaning correctly, even though they did make some difference compared with the sentence-level model.

6.4 *MT on Chinese Pinyin*

Out of the four MT models we used, two of them offer the Chinese pronunciation in Roman alphabets, i.e., Google and Bing, and this is termed "拼音 (Pīn yīn)". The character "拼 (pin)" indicates writing, and "音 (yin)" indicates sounds/reading. We discuss the issues when MT models offer Pinyin for Chinese characters, which are sometimes related to MWEs.

The Bing Translator just kept each Chinese character's Pinyin separated one by one, without any morphological segmentation. GoogleMT's output of Pinyin tried to perform Chinese word segmentation, i.e., concatenate several characters' Pinyin together if they belong to the same word. However, there are several apparent errors in GoogleMT' Pinyin.

1. the Pinyin is wrong itself (totally wrong, or wrong choice in disambiguation situation);
2. the tone is wrong;
3. the word segmentation is wrong which misleads the understanding of the sentence.

In Fig. 19, we listed these three types of error with examples. The first error type was presented in two different situations, with one due to a wrong choice facing ambiguous pronunciation, and the other being totally wrong.

The characters "希律王 (Xī lǜ **wáng**, Herodian)" in the sample sentence with the named entity issue was wrongly annotated as "Xī lǜ **wán**" by GoogleMT, which belongs to a totally wrong situation. [17] The second sentence of this error, "去地精 (qù dì jīng, *de-gnoming*)" was wrongly annotated as "qù dejīng" which is due to the character "地 (dì)" which has different pronunciations[18].

[16] We used sentence level translation outputs of these MT engines since our corpus contains 750 independent sentences that are extracted from the original larger context. However, to have a view of document-level MT performance, we tested some example sentences with context using document-level MT which option is available on the interface of such engines.

[17] Furthermore, the entity "押沙龙 (yā shā lóng, Absalom)" was wrongly segmented into two pieces "押 (yā, bet)" and "沙龙 (shālóng, salon)". This leads to mis-understanding.

[18] https://baike.baidu.com/item/地/34380 and https://en.wiktionary.org/wiki/地

Pinyin Wrong: 100% wrong, Ambiguous wrong	Chiense	**希律王**时代的遗物就是所有的遗物，都应该扭曲成柱状，随着时间的推移经过精心加工，押沙龙的坟墓带有球形的屋顶和奇特的漏斗。(**xī lǜ wáng …** yā shā lóng …) 一知道去地精的事在进行，他们就冲上去观看。 (qù dì jīng)
	Google Pinyin	**xī lǜ wán** shídài de yíwù jiùshì suǒyǒu de yíwù, dōu yīnggāi niǔqū chéng zhùzhuàng, suízhe shíjiān de tuīyí jīngguò jīngxīn jiāgōng, yā shālóng de fénmù dài yǒu qiúxíng de wūdǐng hé qítè de lòudǒu. yī zhīdào qù dejīng de shì zài jìnxíng, tāmen jiù chōng shàngqù guānkàn.
	Ref.	**xī lǜ wáng** shídài de yíwù jiùshì suǒyǒu de yíwù, dōu yīnggāi niǔqū chéng zhùzhuàng, suízhe shíjiān de tuīyí jīngguò jīngxīn jiāgōng, yāshālóng de fénmù dài yǒu qiúxíng de wūdǐng hé qítè de lòudǒu. yī zhīdào qùdìjīng de shì zài jìnxíng, tāmen jiù chōng shàngqù guānkàn.
Tone Wrong	Chiense	椅子很舒服，啤酒已经微微让他上了头。 (wēi wēi) 然后有人说，现在用不了多久，**俄**罗斯人就会让阿拉法特下台。(**é** luó sī)
	Google	yǐzi hěn shūfú, píjiǔ yǐjīng wéiwéi ràng tā shàngle tóu. ránhòu yǒurén shuō, xiànzài yòng bùliǎo duōjiǔ, **è**luósī rén jiù huì ràng ā lā fǎ tè xià tái.
	Ref.	yǐzi hěn shūfú, píjiǔ yǐjīng wēiwēi ràng tā shàngle tóu. ránhòu yǒurén shuō, xiànzài yòng bùliǎo duōjiǔ, **é**luósī rén jiù huì ràng ā lā fǎ tè xià tái.
Segmentation Wrong	Chiense	在72街和**麦迪逊**大道的拐角处，他招手示停了一辆在72街和麦迪逊大道的拐角处，他招手示停了一辆出租车。(**mài dí xùn** … chū zū chē) 在72街和**麦迪逊**大道的拐角处，他招手示停了一辆出租车。(mài dí xùn … guǎi jiǎo … zhāo shǒu shì tíng le)
	Google	zài 72 jiē hé **màidí xùn** dà dào de guǎi jiǎo chù, tā zhāo shǒu shì tíngle yī liàng chūzū chē.
	Ref.	zài 72 jiē hé **màidíxùn** dà dào de guǎijiǎo chù, tā zhāoshǒushìtíng le yī liàng chūzūchē.

Fig. 19 General MT: Pinyin on MWEs

Due to the model failing to acquire the meaning of this term, it chose an incorrect pronunciation.

As another example, the reference sentence from the "common sense" issue in Fig. 19, the name entity "麦迪逊(mài dí xùn, Madison)" was wrongly segmented as "麦迪(màidí)" and "逊(xùn)". These two examples also reflect that the current MT models have issue in recognizing or processing *foreign named entities*. In addition, the vMWE "招手示停了(zhāo shǒu shì tíng le, waved down)" was incorrectly segmented as "招 | 手 | 示 | 停了(zhāo shǒu shì tíngle)" which should be "招手示停 | 了(zhāoshǒushìtíng le)". This illustrates the issue in accuracy level of automatic recognition of vMWE. Furthermore, "拐角(guǎi jiǎo, corner)" and "出租车(chū zū chē, taxi)" should be grouped together as one concept respectively.

6.5 *Simplified vs Traditional Chinese MT*

Traditional Chinese characters have rich linguistic knowledge and are naturally evolved from pictographs via thousands of years of civilization, while simplified Chinese characters are from recent history with less than one hundred years of usage in mainland China. We think it is valuable to offer both these corpora, especially the traditional Chinese characters which are still being used by Taiwan, Hong Kong, Macau, and other regions, such as Kanji characters (made up of traditional Chinese) in Japan.

However, there are some issues in the mismatch translation, for instance, some named entities are translated in different ways by simplified or traditional Chinese by mainland China, Hong Kong or Taiwan with examples in Fig. 20.

In addition to writing system variations, there are different dialects in Chinese, including Mandarin from the Beijing area which is the official language, and Cantonese which is very popular and widely spoken in the Southern part of China including Guangdong province, Hong Kong, and Macau. Cantonese is originally from Guangdong province whose capital city called "Guangzhou" was also known as "Canton" in history.

Out of the four MT models, Bing and Baidu offer both traditional Chinese and Cantonese translation outputs. The difference is that Bing offers Cantonese in traditional Chinese characters, while Baidu offers it in simplified characters. In the example sentence in Fig. 20, regarding the translation of <The Sound and the Fury>, there are at least two issues with the Bing Translator at this point: firstly, when it translates into simplified Chinese, it actually uses the named entities from HK/TW translation < 声音与愤怒, shēng yīn yǔ fèn nù>, just the simplified character of < 聲音與憤怒, shēng yīn yǔ fèn nù>, which is a mess-up. Secondly, when it translates into Cantonese in traditional Chinese, it does not translate the term < 聲音與憤怒, shēng yīn yǔ fèn nù> correctly by dropping < 音, yīn>, in addition to that the overall sentence needs to be improved. Baidu's translation of the Cantonese name < 嘈吵与骚动, cáo chǎo yǔ sāo dòng> is incorrectly replacing "聲音 (shēng yīn, sound)" with "嘈吵 (cáo chǎo, noise)" and "憤怒 (fèn nù)" with "骚动 (sāo dòng)" which is partly from the Mandarin translation.

Regarding, the translation from DeepL, < 声色犬马, shēng sè quǎn mǎ> is a 1974 Chinese movie from HK, which has nothing to do with the title <The Sound and the Fury> in the source sentence.

This literature translation issue reflects the difference between mainland China and other Chinese-speaking regions (Hong Kong, Macau, and Taiwan), also simplified vs. traditional Chinese characters. We need to make the corresponding changes and clarification when we prepare a traditional Chinese corpus for AlphaMWE.

7 Conclusions and Future Work

In this paper, we presented the construction of a multilingual parallel corpus, AlphaMWE, with vMWEs as pioneer annotations by native speakers of the corresponding languages. We plan to extend our annotation into broader MWE categories that are non-verb ones, as well as some terms that might not count as MWEs but present challenges to MT systems, some of which were listed in the MT issues and Discussion section (e.g. "de-gnoming"). We described the procedure of MT system selection, human post-editing, and annotation, compared different state-of-the-art MT models, and classified the MT errors from vMWEs-related sentence/context translations. We characterised translation errors into different categories to help MT research to focus on one or more of them to improve the performance of MT. For instance, how to integrate bilingual terminologies or dictionaries including paraphrases and synonyms to improve metaphorical and idiomatic phrase translation; or is it possible to de-

Source	Reading The Sound and the Fury last night, I came upon words in Compson's thought that belonged to E E Cummings and the thirties, not to the year 1910.
DeepL	2sim: 昨晚读《声色犬马》，我在康普森的思想中看到了属于E-E-康明斯和三十年代的文字，而不是1910年的文字。(shēng sè quǎn mǎ) 2tra/2canto: not available
Google	2sim: 昨晚阅读《喧Sound与骚动》时，我发现康普森思想中的词语属于卡明斯和三十年代，而不是1910年。(xuān Sound yǔ sāo dòng) 2tra: 昨晚閱讀《喧Sound與騷動》時，我發現康普森思想中的詞語屬於卡明斯和三十年代，而不是1910年。(xuān Sound yǔ sāo dòng) 2canto: not available
Bing	2sim: 昨晚读了《声音与愤怒》时， 我在康普森的念头里发现了属于 E E 卡明斯和三十年代的话， 而不是 1910 年。(shēng yīn yǔ fèn nù) 2tra: 昨晚讀了《聲音與憤怒》時, 我在康普森的念頭里發現了屬於 E E 卡明斯和三十年代的話, 而不是 1910 年。(shēng yīn yǔ fèn nù) 2canto: 琴晚讀咗《聲與憤怒》時,我喺康普森嘅念頭度發現咗屬於E E卡明斯同三十年代嘅話,而唔係1910年。(shēng yīn yǔ fèn nù)
Baidu	2sim: 昨晚读《喧哗与骚动》时，我在康普森的思想里发现了一些属于E E·卡明斯和三十年代的词，而不是1910年的词。(xuān huá yǔ sāo dòng) 2tra: 昨晚讀《喧嘩與騷動》時，我在康普森的思想裏發現了一些屬於E E·卡明斯和三十年代的詞，而不是1910年的詞。(xuān huá yǔ sāo dòng) 2canto: 寻晚睇《嘈吵与骚动》嘅时候，我喺康普森嘅思想度发现咗一啲系属於E E·卡明斯同三十年代嘅字，而唔系1910年嘅字。(cáo chǎo yǔ sāo dòng)
Ref.	2sim: 昨晚读《喧哗与骚动》，我在康普森的思想中间或看到了属于E-E-康明斯和三十年代的文字，而不是1910年的。(xuān huá yǔ sāo dòng) 2tra: 昨晚讀《喧嘩與騷動》，我在康普森的思想中間或看到了屬於E-E-康明斯和三十年代的文字，而不是1910年的。(xuān huá yǔ sāo dòng) 2canto: 琴晚睇《聲音與憤怒》果陣時,我發現喺康普森嘅念頭度，係屬於E E卡明斯同三十年代嘅說話， 而唔係1910年。(shēng yīn yǔ fèn nù)

Fig. 20 General MT: Mandarin vs Cantonese, Simplified vs Traditional Chinese (2sim/tra/canto: English to simplified/traditional (Mandarin), and to Cantonese)

sign novel neural network structures to incorporate the MWE compositionality as part of the MT learning stage?

We performed the same process as described here for English→Chinese, English→German, English→Polish, English→Italian, and English→Arabic and similarly categorised the MT issues when handling MWEs. We used English → Arabic corpus as a case study to quantify the error types and ratios using human expert-based post-editing metric HOPE. We also included a future discussion session on corpus development, translation issues on named entities, Chinese Pinyin, Simplified vs Traditional Chinese, etc. where they are related to the processing of MWEs. In future work, we will conduct a more detailed analysis of the English-Italian corpus, in addition to English-German/Polish/Arabic.

We name our corpus AlphaMWE to indicate that we will continue to maintain the developed corpora which are publicly available and extend them into other possible language pairs, e.g. the currently involved Spanish and French under-development. We also plan to extend the annotated MWE genres beyond the vMWEs defined in the PARSEME shared task. There are some limitations on our corpus creation. For instance, even though we have carried out strict quality control on post-editing phase spacially on idiomatic MWEs translation, it might still bias the evaluation towards the MT system that we used to produce the initial raw candidate sentences when this corpus is used for MT system testing and evaluation.

Our AlphaMWE multilingual corpora are available as open access at https://github.com/aaronlifenghan/AlphaMWE.

Author Contributions

LF carried out the initial work on multilingual resource creation of MWEs and co-coordinated the corpus annotation work including the annotation guidelines and recruiting native speakers for annotation. NHM and MR carried out Arabic corpus annotation and quantitative and qualitative evaluation using the HOPE metric. GJ and AS co-supervised the early work on AlphaMWE during LH's PhD and revised the initial manuscript. GN co-supervised the follow-up work on Arabic-MWE [21] and approved this extended manuscript.

Acknowledgements We thank the anonymous reviewers for their valuable comments and suggestions on improving this article. We are very grateful to each target language's post-editing and annotation contributors. ZH: Pan Pan, Qinyuan Li, Ning Jiang; PL: Sonia Ramotowska, Tereska Flera; DE: Daniela Gierschek, Vanessa Smolik, Gültekin Cakir; IT: Gabriella Guagliardo, Paolo Bolzoni. We are thankful for the helpful discussion on English MWEs with Lorin, Roise, and Eoin, and on Arabic with Haifa. The extra examples on English-Polish MT with MWEs in the Appendix are mostly from SR. This work was partially funded by The ADAPT Centre for Digital Content Technology under the SFI Research Centres Programme (Grant 13/RC/2106). The input of Alan Smeaton is partially funded by SFI under grant number SFI/12/RC/2289 (Insight Centre). LH and GN were part-funded by grant EP/V047949/1 "Integrating hospital outpatient letters into the healthcare data space" (funder: UKRI/EPSRC). LH is currently funded by the 4D Picture Project (https://4dpicture.eu/).

Appendix

Extra examples on EN-PL MT issues including MWE-related

We list more interesting examples from English-to-Polish translation, including social/literature-unaware ambiguity, super sense error, context-unaware error, metaphor translation error, abstract phrase MT error, and gender related MT errors.

References

1. Baldwin T, Kim SN (2010) Multiword expressions. In: Handbook of Natural Language Processing, Second Edition, Chapman and Hall, pp 267–292, URL http://www.crcnetbase.com/doi/abs/10.1201/9781420085938-c12
2. Bawden R, Bilinski E, Lavergne T, Rosset S (2020) Diabla: a corpus of bilingual spontaneous written dialogues for machine translation. Language Resources and Evaluation DOI 10.1007/s10579-020-09514-4, URL https://doi.org/10.1007/s10579-020-09514-4

Social/Literature-Unaware Ambiguity	
English	Celebrities, fans and even CNN reporters are dressing up like "Stranger Things" characters this holiday.
MT	Gwiazdy, kibice, a nawet dziennikarze CNN przebierają się za bohaterów "obcych rzeczy" w to święto.
Post-edited	Gwiazdy, kibice, a nawet dziennikarze CNN przebierają się za bohaterów "Stranger Things" w to święto. [sourceVMWE: dressing up][targetVMWE: przebierają się]
The title of the serial "Stranger Things" should not be translated. In Polish we use the original English title.	
English	The moment they know the de-gnoming's going on they storm up to have a look.
MT	W momencie, gdy wiedzą, że zbliża się de-gnoming, burzą się, by rzucić okiem.
Post-edited	W momencie, gdy wiedzą, że zbliża się odskrzatowienie, wzburzają się, by rzucić okiem. [sourceVMWE:going on][targetVMWE: zbliża się]
In this example, “de-gnoming” shall be translated but MT kept it as foreign words in Polish.	

Fig. 21 Social/Literature-Unaware Ambiguity Example in Polish

Super Sense Error Example	
English	He was bearing down on Harry like a great bulldog, all his teeth bared. Original translation: Nosił Harry'ego jak wielkiego buldoga, wszystkie jego zęby wyszczerbiły się.
MT	Nosił Harry'ego jak wielkiego buldoga, wszystkie jego zęby wyszczerbiły się.
Post-edited	Zbliżył się do Harry’ego jak wielki buldog, wyszczerzył wszystkie zęby. Alternatively: Naskoczył na Harry’ego jak wielki buldog, wyszczerzył wszystkie zęby. [sourceVMWE: bearing down (on)][targetVMWE: zbliżył się (do)]
The original translation “Nosił Harry'ego jak wielkiego buldoga” means “he carried Harry like a great bulldog”. We revised it to “moved to Harry” e.g. “Zbliżył się do”. “Naskoczył na” might be even better, because it has this threatening component and it means “move to somebody in the threatening way”.	

Fig. 22 Super Sense Error Example in Polish

Context Unaware MT Errors Example	
English	But it did not give me the time of day.
MT	Ale nie dało mi to pory dnia. (literal word by word translation, in sense of hour)
PE-1: (idiom)	Candidate-1: Ale nie zwrócił na mnie uwagi. (meaning: He did not pay attention to me.) Candidate-2: Był nieprzyjacielski w stosunku do mnie. (meaning: He was unfriendly to.)
PE-2: context	Candidate-1: Ale to nie moja bajka. (Polish idiom “but it is not my fairy tale”, which means “This is not something for me”, like “not my cup of tea”). [sourceVMWE: give (me) the time of day][targetVMWE: nie (moja) bajka] Candidate-2: Ale nie kręciło mnie to. (Polish idiom, which means “This is not something for me”). [sourceVMWE: give (me) the time of day][targetVMWE: nie kręci (mnie)]
Post-editing-1 is based the sentence level translation focusing on the idiom underlined “not give someone the time of day”; PE-2 is based on the context level information.	

Fig. 23 Context-Unaware MT Error Example in Polish

3. Bouamor D, Semmar N, Zweigenbaum P (2012) Identifying bilingual multi-word expressions for statistical machine translation. In: Conference on Language Resources and Evaluation
4. Chowdhary V, Greenwood S (2017) Emt: End to end model training for msr machine translation. In: Proceedings of the 1st Work-

Metaphor MT Errors Example in Polish	
English	The what? Auster laughed, and in that laugh everything was suddenly blown to bits.
MT	Co? Auster się śmiał, a w tym śmiechu wszystko było nagle rozdmuchiwane na kawałki.
Post-edited	Co? Auster się śmiał, a w tym śmiechu wszystko rozbiło się w drobny mak. [sourceVMWE: blown to bites][targetVMWE: rozbiło się w drobny mak]

Abstract Phrase MT Errors Example in Polish	
English	Quinn had his doubts, but this was all he had to go on, his only bridge to the present.
MT	Quinn miał swoje wątpliwości, ale to było wszystko, co miał do zrobienia, jego jedyny most do teraźniejszości.
Post-edited	Quinn miał swoje wątpliwości, ale to było wszystko na czym mógł się oprzeć, jego jedyny most do teraźniejszości. [sourceVMWE: had … doubts; had to go on][targetVMWE: miał … wątpliwości; na czym mógł się oprzeć]

Fig. 24 Metaphor and Abstract Phrase MT Error Examples in Polish

Gender Related MT Errors Example in Polish	
English	Seeing another "Columbia" incident can be quite frightening (I woke up to it early in the morning when I was in Texas).
MT	Widzenie kolejnego incydentu z "Columbią" może być dość przerażające (obudziłem się do niego wcześnie rano, gdy byłem w Teksasie).
Post-edited	Zobaczenie kolejnego incydentu "Kolumbia" może być dość przerażające (uświadomiłem to sobie wcześnie rano, gdy byłem w Teksasie). [sourceVMWE: woke up][targetVMWE: uświadomiłem … sobie]
The term “uświadomiłem to sobie” is a male form, female is “uświadomiłam to sobie”. Here the translation is correct, but there is a general issue that the male form is often a default form e.g. in Google translator. It is important for past tense. In future tense it depends. Also here “woke up to” means “realized” not just “be awake, stop sleeping”.	
English	I have a friend that has to get rid of one of her cats because of allergies, he is the youngest at 3 years old black, long hair, incredibly friendly.
MT	Mam przyjaciela, który musi pozbyć się jednego z jej kotów z powodu alergii, jest najmłodszy w wieku 3 lat czarny, długie włosy, niesamowicie przyjazny. (male form friend)
Post-edited	Mam przyjaciółkę, która musi pozbyć się jednego ze swoich kotów z powodu alergii; jest najmłodszy w wieku 3 lat, długa sierść, niesamowicie przyjazny. (female form friend)

Fig. 25 Gender Related MT Error Example in Polish

shop on Data Management for End-to-End Machine Learning, URL https://www.microsoft.com/en-us/research/publication/emt-end-end-model-training-msr-machine-translation/

5. Constant M, Eryiğit G, Monti J, van der Plas L, Ramisch C, Rosner M, Todirascu A (2017) Survey: Multiword expression processing: A Survey. Computational Linguistics 43(4):837–892, DOI 10.1162/COLI_a_00302, URL https://www.aclweb.org/anthology/J17-4005
6. Gladkoff S, Han L (2021) HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. CoRR abs/2112.13833, URL https://arxiv.org/abs/2112.13833, 2112.13833
7. Gladkoff S, Sorokina I, Han L, Alekseeva A (2022) Measuring uncertainty in translation quality evaluation (TQE). In: Calzolari N, Béchet

F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Odijk J, Piperidis S (eds) Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 1454–1461, URL https://aclanthology.org/2022.lrec-1.156/

8. Han L, Jones G, Smeaton A (2020) AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In: Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, Association for Computational Linguistics, online, pp 44–57, URL https://www.aclweb.org/anthology/2020.mwe-1.6
9. Han L, Jones GJF, Smeaton A (2020) MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 2970–2979, URL https://www.aclweb.org/anthology/2020.lrec-1.363
10. Hassan H, Aue A, Chen C, Chowdhary V, Clark J, Federmann C, Huang X, Junczys-Dowmunt M, Lewis W, Li M, Liu S, Liu T, Luo R, Menezes A, Qin T, Seide F, Tan X, Tian F, Wu L, Wu S, Xia Y, Zhang D, Zhang Z, Zhou M (2018) Achieving human parity on automatic Chinese to English news translation. ArXiv abs/1803.05567
11. Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, Thorat N, Viégas F, Wattenberg M, Corrado G, Hughes M, Dean J (2017) Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics 5:339–351, DOI 10.1162/tacl_a_00065, URL https://www.aclweb.org/anthology/Q17-1024
12. Kato A, Shindo H, Matsumoto Y (2016) Construction of an English dependency corpus incorporating compound function words. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, pp 1667–1671, URL https://aclanthology.org/L16-1263
13. Kato A, Shindo H, Matsumoto Y (2018) Construction of Large-scale English Verbal Multiword Expression Annotated Corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan
14. Lambert P, Banchs RE (2005) Data Inferred Multi-word Expressions for Statistical Machine Translation. In: Proceedings of Machine Translation Summit X, Thailand, pp 396–403, URL https://hal.archives-ouvertes.fr/hal-00705007
15. Läubli S, Castilho S, Neubig G, Sennrich R, Shen Q, Toral A (2020) A Set of Recommendations for Assessing Human-Machine Parity in Language Translation. arXiv e-prints arXiv:2004.01694, 2004.01694
16. Li X, Yan J, Zhang J, Zong C (2019) Neural name translation improves neural machine translation. In: Chen J, Zhang J (eds) Machine Translation, Springer Singapore, Singapore, pp 93–100

17. Macketanz V, Avramidis E, Manakhimova S, Möller S (2021) Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German. In: Proceedings of the Sixth Conference on Machine Translation, Association for Computational Linguistics, Online, pp 1059–1073, URL https://aclanthology.org/2021.wmt-1.115
18. Maldonado A, Han L, Moreau E, Alsulaimani A, Chowdhury KD, Vogel C, Liu Q (2017) Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In: The 13th Workshop on Multiword Expressions @ EACL 2017, ACL
19. Markantonatou S, Ramisch C, Savary A, Vincze V (eds) (2017) Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Association for Computational Linguistics, Valencia, Spain, DOI 10.18653/v1/W17-17, URL https://www.aclweb.org/anthology/W17-1700
20. Markantonatou S, McCrae J, Mitrović J, Tiberius C, Ramisch C, Vaidya A, Osenova P, Savary A (eds) (2020) Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, Association for Computational Linguistics, online, URL https://aclanthology.org/2020.mwe-1.0
21. Mohamed NH, Rassem M, Han L, Nenadic G (2023) Alphamwe-arabic: Arabic edition of multilingual parallel corpora with multiword expression annotations
22. Moreau E, Alsulaimani A, Dublin TC, Maldonado A, Han L, Vogel C, Chowdhury KD (2018) Semantic reranking of crf label sequences for verbal multiword expression identification. Berlin: Language Science Press
23. Pulcini V (2020) English-derived multi-word and phraseological units across languages in the global anglicism database. Textus, English Studies in Italy (1/2020):127–143, DOI 10.7370/97355, URL https://www.rivisteweb.it/doi/10.7370/97355
24. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners
25. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21(1):5485–5551
26. Ramisch C, Cordeiro SR, Savary A, Vincze V, Barbu Mititelu V, Bhatia A, Buljan M, Candito M, Gantar P, Giouli V, Güngör T, Hawwari A, Iñurrieta U, Kovalevskaite J, Krek S, Lichte T, Liebeskind C, Monti J, Parra Escartín C, QasemiZadeh B, Ramisch R, Schneider N, Stoyanova I, Vaidya A, Walsh A (2018) Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In: Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 222–240, URL https://www.aclweb.org/anthology/W18-4925

27. Rikters M, Bojar O (2017) Paying Attention to Multi-Word Expressions in Neural Machine Translation. In: Proceedings of the 16th Machine Translation Summit, Nagoya, Japan
28. Sag IA, Baldwin T, Bond F, Copestake A, Flickinger D (2002) Multiword expressions: A pain in the neck for nlp. In: Gelbukh A (ed) Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–15
29. Savary A, Ramisch C, Cordeiro S, Sangati F, Vincze V, QasemiZadeh B, Candito M, Cap F, Giouli V, Stoyanova I, Doucet A (2017) The PARSEME shared task on automatic identification of verbal multiword expressions. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Valencia, Spain, pp 31–47, DOI 10.18653/v1/W17-1704, URL https://www.aclweb.org/anthology/W17-1704
30. Schneider N, Onuffer S, Kazour N, Danchik E, Mordowanec MT, Conrad H, Smith NA (2014) Comprehensive annotation of multiword expressions in a social web corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), European Languages Resources Association, Reykjavik, Iceland, pp 455–461, URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf
31. Skadina I (2016) Multi-word expressions in English-Latvian machine translation. Baltic J Modern Computing 4:811–825
32. Sun M, Jiang B, Xiong H, He Z, Wu H, Wang H (2019) Baidu neural machine translation systems for WMT19. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Association for Computational Linguistics, Florence, Italy, pp 374–381, DOI 10.18653/v1/W19-5341, URL https://www.aclweb.org/anthology/W19-5341
33. Tan L, Pal S (2014) Manawi: Using multi-word expressions and named entities to improve machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA, pp 201–206, DOI 10.3115/v1/W14-3323, URL https://www.aclweb.org/anthology/W14-3323
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Conference on Neural Information Processing System, pp 6000–6010
35. Vincze V (2012) Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association, Istanbul, Turkey, pp 2381–2388, URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/177_Paper.pdf
36. Vincze V, Csirik J (2010) Hungarian corpus of light verb constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, pp 1110–1118, URL https://www.aclweb.org/anthology/C10-1125

37. Vincze V, Nagy T I, Berend G (2011) Multiword expressions and named entities in the wiki50 corpus. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Association for Computational Linguistics, Hissar, Bulgaria, pp 289–295, URL https://www.aclweb.org/anthology/R11-1040
38. Walsh A, Bonial C, Geeraert K, McCrae JP, Schneider N, Somers C (2018) Constructing an annotated corpus of verbal MWEs for English. In: Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 193–200, URL https://www.aclweb.org/anthology/W18-4921
39. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144, URL http://arxiv.org/abs/1609.08144
40. Zaninello A, Birch A (2020) Multiword expression aware neural machine translation. In: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 3816–3825, URL https://www.aclweb.org/anthology/2020.lrec-1.471