# AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations

Najet Hadj Mohamed [1,2*] Malak Rassem [3*] Lifeng Han [4] Goran Nenadic [4]
[1] University of Tours, France
[2] Arabic Natural Language Processing Research Group, University of Sfax, Tunisia
[3] Institute for Natural Language Processing (IMS), University of Stuttgart, Germany
[4] Healthcare Text Analytics Group, University of Manchester, United Kingdom
* *co-first authors*

**RANLP2023**: Recent Advances in Natural Language Processing
Varna, Bulgaria, Sep 2-8

université de TOURS

Université de Sfax
TUNISIE

Universität Stuttgart

MANCHESTER 1824
The University of Manchester

# Resources:

Link to download our paper

Link to the corpus and data: https://github.com/aaronlifenghan/AlphaMWE

Original alphaMWE corpus presentation: https://youtu.be/KiuF5JdOlLw

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

README.md

The original English source repo (https://gitlab.com/parseme/parseme_corpus_en)

Five portions of the files from 'aa to ae', 150 segments each.

License: since we use the English PARSEME dataset, we adopt the same license as the original dataset, i.e. CC-BY-SA 4.0

If you are interested in including your native languages into AlphaMWE (currently involved: English/Chinese/German/Polish/ working:Arabic/Italian/Spanish), please get in touch. We do think this is a good contribution to various native language processing in machine / AI era, in addition to lexical studies.

# Download multilingual parallel corpora (en, de, zh, pl, ar (working), it (working) | English-German-Chinese-Polish-Italian-Arabic)

# Motivation

- Expanding AlphaMWE parallel corpus
- Low resource language: Dialectal Arabic
- How good is Modern Standard Arabic (MSA) MT output?

# Methods

- HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation (Gladkoff & Han, LREC 2022) https://aclanthology.org/2022.lrec-1.2
- Post-editting MSA by native speakers => cross validation
- Dialectal Arabic: Translation from scratch by native speakers
- - Tunisian Arabic by Najet Hadj Mohamed
- - Egyptian Arabic by Malak Rassem

# On Arabic

The MSA: Modern Standard Arabic

- No capital letters and Not-widely adopted punctuation marks
- Tend to use long and complex sentences with right-to-left writing
- Complex morphology as a Semitic language: e.g. concatenative morphology
- Diacritical marks on top or below symbols to form new words: e.g. new pronunciations and meanings - causing ambiguity for certain texts without diacritical marks
- Flexible word order

=> read our paper for examples?

# On AlphaMWE

Root AlphaMWE-EN: extracted from PARSEME 2018 corpus on vMWE identification and discovery shared task challenge (Walsh et al. 2018; Ramisch et al. 2018)

ADAPT-Ireland Team attended the shared task and ranked one of the best.

Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking
A Maldonado, L Han, E Moreau, et al.
The 13th Workshop on Multiword Expressions @ EACL 2017

PARSEME-EN:

- 832 vMWEs were manually annotated in 7,437 sentences of context.
- Various topics: news, literature, IT documents.

AlphaMWE-EN:

- All 750 sentences extracted that have vMWE labels
- Divided into 5 portions: 150 sentences each x 5 => 750: Cross validation, system fine-tuning, testing, etc.

# On MT System Selection

| | *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 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 حافظات 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 ، ونفض 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 قاسية، وأزاغ 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 وتلقائي. 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 للوهلة 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 الرهائن 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 كأحداث أخرى. |

# On MT System Selection

1) when **SysTran** MT output makes mistakes, the errors are very severe, such as adding context out of the blue, while **GoogleMT's** output still makes some sense when it is wrong.

2) **SysTran** has more correct translations on entities.

To reduce the workload for the professional post-editing step;

To know more about how MT makes mistakes when translating MWEs and verbal idioms

=> We choose GoogleMT as our engine

a) entity errors can be fixed more easily than out-of-the-blue errors;

b) we can get more examples of how MT fails in translating MWE-related content => can be valuable for future research such as on guiding MT development.

# On MT-evaluation Metric: HOPE

Eight designed error types: according to industrial practice.

Different level of error severity scores per error:

2^n = (0, 16):  (0, 1, 2, 4, 8, 16)

Sentence/segment-level classification: good-enough/minor (1-4), major (5+)

We added **two more** error types for this study:

- MMC: missed chance - when the MT output on source MWEs is either wrong semantically or correct translation but without using the corresponding MWEs
- SKP: skipped word - MT system fails to translate certain words that are important

=> check paper for scoring example sentence

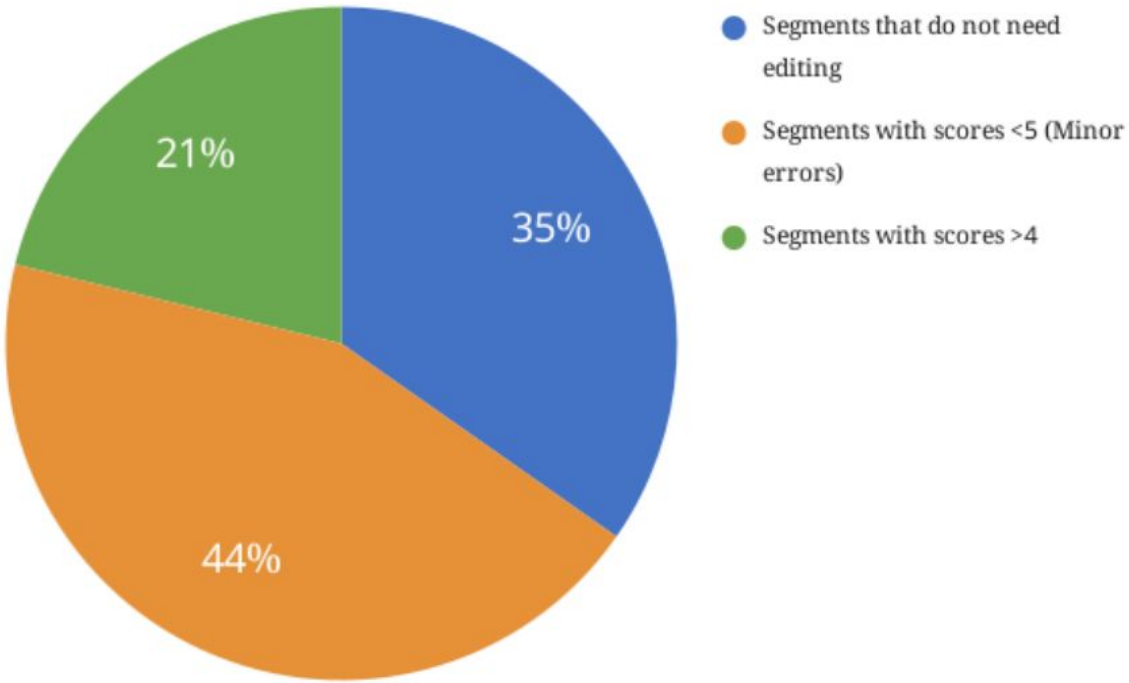
Serge Gladkoff and Lifeng Han. 2022. HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

# On MT-evaluation Metric: HOPE

| Code | Definition | Explanation |
|---|---|---|
| | | |
| **IMP** | Impact | The translation fails to convey the main thought clearly (even if translation may be literally correct, but proper translation should not be literal in target language, or has poor expression of the main thought). |
| **RAM** | Required Adaptation is Missing | Source contains error that has to be corrected, or target market requires substantial adaptation of the source, which translator failed to make. Impact on end user suffers. |
| **TRM** | Terminology | Incorrect terminology, inconsistency of translation of entities (forms, sections, etc.) |
| **UGR** | Ungrammatical | Translation is ungrammatical - needs to be fixed to convey the meaning properly. |
| **MIS** | Mistranslation | Translation distorts the meaning of the source, and presents mistranslation or accuracy error. |
| **STL** | Style | Translation has poor style, but is not necessarily ungrammatical or formally incorrect. |
| **PRF** | Proofreading error | Linguistic error which does not affect accuracy or meaning transfer, but needs to be fixed. |
| **PRN** | Proper Name | A proper name is translated incorrectly. |

# MT Error Analysis: Statistics

Penalty Score Ratios of Each Error Level (0, minor, major) from 150 Segments using HOPE Metric.



# MT Error Analysis: each error type

| Error type | MMC | MIS | STL | TRM | IMP | UGR | PRF | SKP | All | PPS |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Penalty Scores | 76 | 68 | 69 | 39 | 114 | 37 | 46 | 6 | 455 | |
| Ratio out of total segments | 17% | 15% | 15% | 9% | 25% | 8% | 10% | 1% | | 3.03 |

**All error types:**

Mistranslation (MIS)
Style (STL)
Terminology (TRM)
Impact (IMP)
Missing Required Adaptation (RAM)
Ungrammatical (UGR)
Proofreading Error (PRF)
Proper Name (PRN)
*MWE Missed Chance (MMC)*
*Skipped Word (SKP)*

- Highest: IMP, **MMC** (we added)
- Followed by: MIS, STL
- SKP: around 1% of all errors
- All error scores: 455
- Penalty Per-Segment: 3.03

# Outcomes

An Arabic corpus with MWEs annotated including three subsets:

- MSA corpus yielded 2,700 tokens
- Tunisian Arabic: 2,495 tokens translated
- Egyptian Arabic: 2,055 tokens translated

MT Error Analysis with Qualitative and Quantitative annotations

- HOPE metric Excel sheet - can be used by linguists for further study

# Example of Generated Annotation using HOPE

| SRC | SRC MWE | Google Translate | MMC | MIS | STL | TRM | IMP | UGR | PRF | SKP | SEGS | Human Gold Standard | Target MWE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Also, you can set the line style of borders. For example, you can specify that a border consists of dashes or dots. | sourceVMWE: consists of | أيضًا ، يمكنك تعيين نمط خط الحدود. على سبيل المثال ، يمكنك تحديد أن الحد يتكون من شرطات أو نقاط. | | | 1 | | | | | | 1 | أيضًا ، يمكنك ظبط نمط خط الحدود. على سبيل المثال ، يمكنك تحديد أن الحد يتكون من شرطات أو نقاط. | يتكون من |
| The type of interval you can specify for a field depends on the data type of the field. | sourceVMWE:depends on | يعتمد نوع الفاصل الزمني الذي يمكنك تحديده للحقل على نوع بيانات الحقل. | | | | 1 | | | | | 1 | يعتمد نوع الفاصل الزمني الذي يمكنك تحديده للخانة على نوع بيانات الخانة. | يعتمد |
| 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. | sourceVMWE:taken over | كان هذا مكانًا كئيبًا بشكل عام ، مليئًا بالغبار والناس ليس لديهم مكان يذهبون إليه ، ولكن الآن ، مع ساعة الذروة بكامل قوتها ، استولى عليها الرجال والنساء مع حقائبهم وكتبهم وصحفهم. | | | 4 | | 4 | | | | 8 | كان هذا مكانًا كئيبًا بشكل عام ، مليئًا بالغبار والناس التي ليس لديهم مكان يذهبون إليه ، ولكن الآن ، مع ساعة الذروة بكامل قوتها ، استولى عليها الرجال والنساء مع حقائبهم وكتبهم وصحفهم. | استولى علي |
| 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. | sourceVMWE:made it; gave … s | لقد وصل إلى الفقرة الثالثة أو الرابعة عندما استدار الرجل ببطء تجاهه ، وأعطاه نظرة شريرة ، ونفض الورقة بعيدًا عن الأنظار. | | | | | 1 | | | | 1 | لقد وصل إلى الفقرة الثالثة أو الرابعة عندما استدار الرجل ببطء تجاهه ، وأعطاه نظرة شرسة ، ونفض الورقة بعيدًا عن الأنظار. | وصل |
| She was, however, reading a book, a paperback with a lurid cover, and Quinn leaned over ever so slightly to his right to catch a glimpse of the title. | sourceVMWE: catch…glimpse | كانت ، مع ذلك ، تقرأ كتابًا ، غلافًا ورقيًا بغلاف فاضح ، وانحنى كوين قليلاً إلى حقه في إلقاء نظرة على العنوان. | | 8 | | | 4 | | | | 12 | كانت ، مع ذلك ، تقرأ كتابًا ، غلافًا ورقيًا بغلاف فاضح ، وانحنى كوين قليلاً جداً إلى يمينه لإلقاء نظرة على العنوان. | إلقاء نظرة |
| But now that the scene was taking place, he felt quite disappointed, even angry. | sourceVMWE:taking place | ولكن الآن بعد أن بدأ المشهد ، شعر بخيبة أمل كبيرة ، وحتى غاضب. | | | | | | 8 | 1 | | 9 | ولكن الآن ، بعد أن بدأ المشهد ، شعر بخيبة أمل كبيرة ، وحتى ببعض الغضب. | بدأ |

# Discussion and Future Work

- Completion of AlphaMWE-Arabic corpus: two more files out of five.
- Inter-annotator agreement: is it possible to calculate for this task?
- More Arabic dialects => have interests? get in touch?

Link to download our paper

Link to the corpus and data: https://github.com/aaronlifenghan/AlphaMWE

Original alphaMWE corpus presentation: https://youtu.be/KiuF5JdOlLw

# ألفا

# MWE

## اللغة العربية

# AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations



Authors Najet Hadj Mohamed, Malak Rassem, Lifeng Han, Goran Nenadic

Publication date 2023/3/17

Description Multiword Expressions (MWEs) have been a bottleneck for Natural Language Understanding (NLU) and Natural Language Generation (NLG) tasks due to their idiomaticity, ambiguity, and non-compositionality. Bilingual parallel corpora introducing MWE annotations are very scarce which set another challenge for current Natural Language Processing (NLP) systems, especially in a multilingual setting. This work presents AlphaMWE-Arabic, an Arabic edition of the AlphaMWE parallel corpus with MWE annotations. We introduce how we created this corpus including machine translation (MT), post-editing, and annotations for both standard and dialectal varieties, ie Tunisian and Egyptian Arabic. We analyse the MT errors when they meet MWEs-related content, both quantitatively using the human-inthe-loop metric HOPE and qualitatively. We report the current state-of-the-art MT systems are far from reaching human parity performances. We expect our bilingual English-Arabic corpus will be an asset for multilingual research on MWEs such as translation and localisation, as well as for monolingual settings including the study of Arabic-specific lexicography and phrasal verbs on MWEs. Our corpus and experimental data are available at https://github.com/aaronlifenghan/AlphaMWE.