Guidelines for Egyptian Arabic-English Word Alignment

Version 2.0 – 04/10/2014

Xuansong Li, xuansong@ldc.upenn.edu
With contributions from:
Stephanie Strassel, strassel@ldc.upenn.edu
Safa Ismael, Petra Brayo, Dalal Zakhary

Linguistic Data Consortium

Table of Contents

1. Introduction

The word alignment task consists of finding correspondences between words, phrases or groups of words in a set of parallel texts. The resulting data can be used as gold standard training and testing data for machine translation. With references to Blinker, ARCADE project guidelines and GALE Arabic word alignment guidelines, these guidelines are especially designed to suit the task of Egyptian Arabic-English word alignment. A visualized tool is developed by LDC to facilitate the task.

1.1 Overview of the Guidelines

The data used for word alignment is first presented in Section 2. In the Section 3, the tasks are specified, and the conventions adopted in the guidelines are explained in greater detail. In Section 4, the general strategies of annotation are addressed to deal with universal language features in word alignment. More detailed specifications and rules are illustrated with examples in Section 5. Section 6 and Section 7 are devoted to approaches toward distinctive features of Arabic and Egyptian Arabic languages.

2 Data

Source data genre type is discussion forum. Egyptian Arabic words are tokenized automatically by ATB team at LDC, and we used Egyptian treebank leaf tokens for alignments. Each Egyptian Arabic word is treated as a separate token; English words in Arabic texts are tokenized according to Penn English Treebank standards; punctuation is separated; clitics are separated from the stem words. In Egyptian Arabic word alignment, all hyphens are treated as separate tokens except when used in transcription to indicate a partial word.

English tokenization follows the same guidelines used in Penn English Treebank: split words by white spaces, separate punctuations from the preceding/following words, apostrophes S (‘s) are treated as separate tokens. Penn English Treebank treats most hyphens as separate tokens but some as part of words.

3 Tasks and Conventions

3.1 Tasks

a) Link words or phrases in the source language (i.e. Arabic) to those in the target language (i.e. English).
b) Make judgments on translated and/or not-translated elements in the source and target languages.
c) Attach unmatched words to their related parts according to attachment rules.

d) Reject the alignment using the “Reject Segment” button for blank sentences, unmatched sentences, half translated sentences and pure English sentences on both sides.
e) Add a comment if any potential problems are spotted (optional).

3.2 Conventions

To better understand the word alignment annotation task and to facilitate the compilation of the guidelines, the following terms and symbols are employed.

a) Translated, not-translated, correct, incorrect are labels that appear in the tool for links and markups.
b) Brackets <> are used to indicate equivalence between the source language and the target language. E.g. president <> رئيس
c) Parenthesis ( ) are used to indicate emptiness or omission, E.g. `()<> the` This indicates “the” appears in the translation but no counterpart in the source language.
d) Colors, brackets, asterisks and plus signs in the guidelines indicate corresponding links and markups: brackets <> represent alignment, asterisks (*) represent attachment of unaligned words, plus or minus signs (+, -) represent the split of tokens. Green highlights represent correct alignment links. Blue highlights represent “not translated and correct” words. Red highlights represent “not translated and incorrect” words. Yellow highlights represent incorrect alignment links. Italics are used for typos in Egyptian Arabic source, and bold letters for tokenization errors.

4 Concepts and General Approaches

For efficiency and accuracy, it is important to emphasize some general alignment strategies.

- The annotator should glance over both the source sentence and target sentence in the two windows on the right side of the tool interface to get to know the sentence and its context and to make sure the translation is valid.
- Beginning with the source sentence, the annotator should first link all the content words before moving on to function word links.
- With all the equivalent links mapped, the left-over words or phrases can be either aligned to other parts or marked as “*not-translated”*.
- Most of the links are “translated correct” links; “not-translated correct” links are used for purely functional or grammatical word insertions due to language idiosyncratic features or for contextual insertions for the sake of effective communication and discourse coherence.
- All words in both source and target languages should either be linked or marked. No word should be left unattended. An annotator can reject a sentence if s/he thinks that sentence is not suitable for alignment.

4.1 Translated versus Not-translated

4.1.1 Translated

When translating from one language to another, it is common to see multiple translation versions for a particular word or a sentence in the source language. The different versions may all convey the same meaning, i.e., they are semantically equivalent but the difference may occur as a result of a word choice or style. Thus, each of these versions is a correct translation of the source words or phrases. Therefore, all these versions are considered to be “translated” and “correct”. For mapping, an obvious lexical item or items can be found in the source and translation.

One source word, for example, can be translated in different ways (synonyms).

Example:

عرض المدير رفض <> (version 1) **He* refused the manager’s offer.**
(version 2) **He* declined the manager’s offer.**
(version 3) **He* said no to the manager’s offer.**

A pair is also treated as “translated” and “correct” even if there is a change in part-of-speech or sentence structure.

An example to show change of part of speech:

اكيد ل- كل واحد مفهوم- ه- الخاص
Certainly everyone has his own concept

يبقى شوية شوية نتعامل مع الامور الاخرى
So we need to deal bit by bit with the other issues

عمر- ك- شلت شوالات على ضهر- ك- ؟
Have you ever carried sacks on your back?

An example to show the surface structure change:

يكفى وجود صلة ارحام و- انساب بين- نا و- بين- هم
It* is sufficient that there is kinship relations between them and us.

There are two types of links in terms of “translated” type:

1) Translated and incorrect

If a word or phrase is translated in a wrong way -- either semantically or grammatically -- they are treated as translated but incorrect. Typos or grammatical errors in the target language can be treated as “translated” and “incorrect”.

An example to show semantically incorrectly translated:

ثانياً ب- -احب بلد- -ي جداً و- -نفس -ي اشوف- -ها احسن بلد في الدنيا
Once again, I love my country and I wish to see it the best country in the world.

An example to show a non-grammatical translation:
يا- -ساتر ده انت ملل
Geeez your so boring

آل ايه بقى دولم بتوع الخنوع
So they say now that those is the submissiveness lot

An example to show a typo in translation.
هوه فعلاً خجول و- -كل ما يشوف- -ني ما يقول- -ل-* -ي- -ش غير يا- -ست هانم
He is indeed shy and whenever he sees me he would only call be ma'am

2) Translated and correct

Translated and correct links are the default and are the most frequent type. The meaning is conveyed properly and they are grammatically correct. The links where direct equivalences of both form and content could be spotted are the easiest ones to detect.

An example to show obvious and direct equivalence between source and translation:
و- -القطط الغلبانة اللي بتاكل من الزبالة دي
And the poor cats that eat from this trash.

A typo in the source side can still make a correct link to the translation when the translation is right.

An example to show a typo in the source:

ايوا زي القرد و- -الاسد في الغابة ياما سمعنا حد/ويت في- -ها حيل.
Yes like the monkey and the lion in the jungle, O how often we heard tales that had tricks.

4.1.2 Not-translated

A word is marked as "not-translated" when it is both semantically and lexically missing from the English translation or when the lexical representation of the word is missing from the translation but the semantic meaning is neither lost nor added. In such cases, extraneous words could be found. Thus two cases are recognized for "not-translated" markups:

1) not-translated and incorrect:

This kind of markup is proper when both the word form and the information are lost.

An example to show both word form and the information are lost in translation:

ف- -طول عمر- -نا بنشوف ان ناس احسن من- -ه مليون مرة و- بيتهاجموا ب- -كل حدة و- -شدة بينقبض علي- -هم و- بيتشهر بيهم و- -ب- -سمعتهم و- -دي ضريبة النجاح.

So all our lives we have seen people a million times better than him who are criticized more severely and strongly get arrested.

In translation, we sometimes find the meaning of a word translated partially. In other words, we see partial concept entailment between the source word and translated word. We treat these words as “not-translated” and “incorrect” because partial meaning is lost anyway.

Examples to show word partial entailment in translation:

و- -في الآخر عايز اقول حاجة مهمة اوي

In the final analysis, I would like to mention something very important.

اي حد يشتهر في فترة قليلة او شهرت- -ه تزيد في فترة قليلة من غير سبب مقنع.

When anyone becomes famous in a short period or his fame increases in a short period without a reason.

Topic related extra words, which are usually content words, are treated as “not-translated” and “incorrect” because the topic clue may be found within or beyond a sentence, and such inference may simply rely on reader’s understanding of the discourse focus.

An example to show discourse-level content word insertions:

آثار النوبة ل- -علاء ... و- -الاقصر ل- -جمال

Hosni: The Nubian monuments go to Alaa … and Luxor goes to Gamal.

2) not-translated and correct:

“Not-translated correct” refers to extra words inserted during translation. These words are either grammatically or contextually needed to make the translation fluent. For instance, the words “و or ف” in the following sentences are all extra words needed for Arabic and English to make the sentences grammatical and fluent.

و- -لا يهم- -ك يا- -بابا

Not to worry Dad.

انا مستعد- ف-
I am ready

Extra words can be inserted at the phrase level, the sentence level or a discourse level. For phrase (or local) level insertions, we use an attachment approach to align them to their related words (see Section 5). For other insertion cases not mentioned in section 5, we mark all of them as “not-translated correct” (see Section 6: unmatched/unattached words).

4.2 Minimum-Match

4.2.1 Minimum-Match in Literal Translations

In the guidelines, the principle of a word-for-word alignment is strictly followed, that is, the smallest number of words is preferred.

The following examples to show minimum match (one-to-one translation):

المصريين مختلفين عن باقي العرب
Egyptians are different from the rest of the Arabs

و- اقول ل- الاخ بازوكا الكنتالوب اللي مش عاجب- ك ده ارخص من الخيار با- حبيب- ي-
And I say to brother Bazuka: the cantaloupe that does not appeal to you is cheaper than cucumber, my dear.

4.2.2 Maximum-Match in Idioms and Non-Literal Translations

There are cases where a minimum-match cannot be achieved. In such cases, to ensure a one-way semantic equivalence, annotators will need to choose as many words as necessary for one link. That is, select as many characters as you can to make a semantically independent unit. Such cases include formulaic or frozen expressions, idioms (or near idioms), proverbs and hyphenated words. Other cases include very tricky structures such as verb phrases (verb particles) and words with affixes.

The following examples to show the alignment of idioms, fixed expressions, and proverbs:

جت جت ما- جت- ش عن- ها ما- جت-
If it happens, it happens, if it doesn’t happen, may it never happen.

امسكوا ب- -ايدن- -كم و- -سنان- -كم
Hold on with al you 've got.

يمين يمين شمال شمال
You say right and right he goes, you say left and left he goes

الله يرحم- -ك
Rest in peace

اكفي القدرة علي فم- -ها تطلع البنت ل- -ام- -ها
Like father like son

For proverbs, however, if a word-for-word translation is found, sub-part links would be preferred especially for those directly borrowed from English.

يضرب عصفورين ب+ حجر واحد
Kill two birds with one stone.

القافلة تسير و- -الكلاب تعوي
The dogs bark but the caravan moves on.

We follow the same approach for proper nouns, which are described in Section 5.5.

Non-literal translations -- other than idioms or proverbs -- are also very common. In such cases, drilling down to individual constituents would result in semantically non-equivalent alignments. To avoid such non-equivalent alignments, we would prefer to treat them as whole units.

ل- -ك الله
May God be with you

Hyphenated words in English are treated as one unit if the subparts of the hyphenated words are inseparable. In tokenization, these inseparable words are treated as one word. For those which are not tokenized as one word, we will not break down into subpart alignments.

An example to show hyphenated words are aligned as a whole unit:

الثورة المضادة
Counter-revolution

4.3 Attachment Approach

Extraneous words on the phrasal level are usually attached to their dependency constituents when they are grammatically or contextually required (see Section 5 on attachment rules).

We align extra words to their generating heads in complex situations. Such words can be conjunctions or prepositions.

Examples: the highlighted words should be aligned to their heads**:**

صحيح العنف مرفوض لكن استحالة نقبل ب- -وجود افكار ممكن تدمر عقيدة اولاد- -نا بين- -نا
Really, violence is rejected, but it is impossible to accept idea that* can devastate the ideology of* our children among us.

بس لو كان انحل مش المستشارين اللي عدلوا في-* -ه كان يبقى ده رأي- -هم
But if it has been nullified and not that the advisors who amended it, then this is their opinion.

المسامح كريم
Forgiveness is* noble.

بس انا دايماً بحب اقرا مواضيع- -ك يا* فندم
But I always like to read your articles sir.

5 Alignment and Attachment Rules

5.1 Anaphora (pronouns)

Anaphora are the co-reference of one expression with its antecedent. The links and markups can be done as follows.

Case 1: Pronoun <> Pronoun

ب- -جد دماغ- -ي لفت يا- كيمو
Really you got my head turning, Kimo.

Case 2: Pronoun <> Verb

في اسوأ الاحوال اللي بيعيش هو الفايز.
In the worst cases, the winner is the one who survives.

In Arabic, a subject is sometimes dropped after its first mention whereas it is kept in English. Although omitted, the Arabic subject is understood from the verb. We

attach the pronoun or subject to its referents and consider them as “translated” and “correct”.

The following examples to show the omission of subject:

Case 3: Verb <> Pronoun+Verb

يـاريـت يـسلفـوا -هـ -ل- -نـا مـحتـاجين- -هـ
I wish they* lend him to us. We need him.

يـا- -تـرى بـتفـكر و- -بـتقـول ايـه دلـوقـتي
Can you imagine what she* 's thinking or saying now.

Case 4: Pronoun <> Referent

انـا رأي شبـاب زي دول كـانـوا لـقـوا شغل فـي بـلد- -هـم و- -قـالـوا لا قـبل مـا نـلـوم- -هـم على حـاجة زي دي
My opinion: have youths like these found work in their country and said no Just before we blame the youth for something like this.

Case 5: Referent <> Pronoun

عند وصول هذا الأخير
Upon his arrival

We adopt the same attachment approach for reflexive pronouns (myself, yourself, herself, himself, itself, ourselves, yourselves) and indefinite pronouns (one, some, someone).

5.2 Demonstrative Words

Links involving demonstratives (this, that, these, those, here, there) can be illustrated by the following cases.

Case 1: Demonstratives <> Demonstratives

فـي مـشاكـل كـتـيـرة هـنـا او هـنـاك
There are many problems here and there.

انـا مـالـيش اي مـصلحة فـي نـشر الـكلام ده
I do not get anything back for posting these words.

Case 2: Demonstratives <> Referent (or vice-versa)

هـاتـوا الـمـخرج الايـطـالـي الـلـي اسم- -هـ مـكـافـيـلـلي يـصور- -هـا
Get this Italian director, the one called Machiavelli to film it.

دة راجل يا اهبل يا مجنون
The man is either foolish or crazy

As illustrated in Case 2 above, demonstratives with no translation equivalent are attached to their head word(s).

5.3 Copula BE

Link verb “to be” (am, is, are, be, been, being) qualifies or provides information about the properties of its argument. Many languages do not use the verb "to be" to separate the subject from its object, such as in Arabic, then in link mapping, the extra link verb can be treated as “not-translated” and “correct”.

An example to show “be” is aligned:

كل اللي بيقولوا على- -ه انجازات دي مش انجازات
All these things that they are calling achievements are not achievements.

Examples to show “be” are omitted:

على فكرة هو ظروف- -ه تقريباً زي ظروف- -ي ب- -الظبط
By the way, his circumstances are* almost exactly like mine.

5.4 Proper Noun

A proper noun is used to name a specific item. It is usually a one-of-a-kind item and it always begins with a capital letter.

We follow the basic minimum-match approach in aligning proper nouns, especially when they are compositional (usually one or more specific name plus one or more common nouns).

عبد الرحمن الكواكبي
Abdulrahman Al-Kawakibi

أم كلثوم
Um Kulthoum

سيدي أحمد بن بيلا
Sidi Ahmed Ben Bella

أم ايمن
Om Ayman

For names of people, we align first name and last name separately.

حسني مبارك
Hosni Mubarak

باراك اوباما
Barack Obama

For names of countries and such entities in their abbreviated forms, we treat them as whole units.

البيت الأبيض
The While House

كفر مصيلحة
Kafr Meselha

When the name is not abbreviated and has good one-to-one match, we'll break down to sub-piece alignment.

الولايات المتحدة الأمريكية
The United States of America

Acronyms of translation of proper nouns are treated as whole units.

الناتو
North Atlantic Treaty Organization

أنروا
United Nations Relief and Works Agency

However, if there is no perfect match for the case of abbreviations for proper nouns, then we align them as a whole unit.

يو إن إتش سي آر
UNHCR

5.5 Determiners (articles in English)

Determiners -- particularly articles in these guidelines -- are used before nouns to indicate whether one is referring to a specific object or a particular type. The definite and indefinite articles (a/an/the) are all determiners. Such determiners are attached to their head nouns. In Arabic, however, there are no indefinite articles but there is only one definite article, "الـ". It always comes attached to the noun so we align it together with the noun.

Sentences involving articles can have the following cases:

Case 1: Determiner <> Determiner

ب- -لا هيافة و- -قلة قيمة وسط العالم بقى
With no trivialities and lack of worth in the middle of* the world, then.

Case 2: (none) <> Determiner (or vice-versa)

طيب ما- -نشوف حل تاني بدل ما- -نكون فريسة سهلة ل- التجار و- الحرمية
Ok, why do not we look for another solution instead of being an easy prey for businessmen and thieves

ربنا يستر على الامة الاسلامية لو مباراة بتعمل كدة
Lord, protect the Islamic nation if a tournament is played like this.

It is common to have a determiner in one sentence that is omitted in the translation. We align the extra determiners to their head noun even if the head noun is not adjacent to the determiner.

Case 3: Determiner <> different translations (or vice-versa)

See examples above under Section 5.2 case 2.

It happens occasionally that a determiner in one sentence segment has a non-determiner translation equivalent. In this case, the alignment is created according to the meaning expressed.

5.6 Auxiliary Verbs

An auxiliary or helping verb is a verb giving further semantic information about the main or full verb following it. An auxiliary verb can have many forms. Auxiliary verbs are used to serve such functions as passive, progressive, perfect, modal, or dummy. Whenever the auxiliary verbs appear in both sentences, simply align them. When missing in one sentence, the extra auxiliary verbs will align to the main verb.

An example to show auxiliary used together with negative words:

مع ان وائل غنيم ساعة- -ها ما- -كان- -ش مشهور و- -لا حاجة
Although Wael Ghoneim at that time was not famous or anthing.

مع ان وائل غنيم ساعة- -ها ما- -كان- -ش مشهور و- -لا حاجة
Although Wael Ghoneim at that time was not famous or anthing.

An example to show how extra auxiliary verbs are aligned:

نجم الاخوان و- السلفيين سطع
The star of the Brotherhood and the Salafis was rising

The word "do" may be used for "emphasis"; in this case it should be "not translated correct".

انا بكره الفساد اللي كان موجود في البلد
I do hate the corruption that was all over the country.

5.7 To- Infinitive

In Arabic-English alignment, "to" infinitive is aligned with the verb following it. Occasionally, some words might be inserted between infinitival to and the verb, a case which requires extra attention to make a correct alignment.

An example to show words are inserted between an infinitive "to" and the verb:

نفس- ي مرة اقول احنا عملنا حاجة
I wish once to say that we have done anything

An example to show the reverse in Arabic:

ف- كيف ل- هذا الرجل ان يفرط في فتفوتة من خبز ؟
How then could this man squander a crumb of the bread?

Occasionally, "to" infinitive may find its translation equivalent, where a direct link would be possible.

An example to show "to" is aligned:

واجب على كل مواطن مصري ان يحب بلد- ه
Every Egyptian citizen has the duty to love his country

5.8 Expletives

Here "expletive" mainly refers to syntactic expletives: words that perform a syntactic role but contribute nothing to meaning, such as "it", "there" "here". For cases where an equivalent counterpart can be found, we link them. If without equivalence, the extra "it", "there" or "here" can be marked as "not-translated" and "correct".

An example of an unaligned expletive:

بس الصراحة غالية اوي
But, frankly, it is very expensive.

5.9 Conjunction

5.9.1 Conjunctions with “and”

Four cases of conjunctions are involved under this category:

a) و <> and
b) , <> , and (or ,<>, و)
c) و <> ( ) (or () <> and)
d) و <> ()

For case (a), simply link both sides.

و- انا اللي كنت فاكر- ك مهتم بيا و- بتنور- ني
And I thought that you cared and were enlightening us.

For case (b), link comma to comma, and the “and” to the “and”

ل- كل قمة قاع , و- ل- كل قاع قمة
For every summit there is a bottom , and for every bottom there is a summit.

For case (c), link comma to “and”, and extra conjunctions are marked as “not translated and correct”

الجميلة ب- أهل- ها و- طقس- ها و- آثار- ها
Beautiful with its people , its climate , and its antiquities

For case (d), the extra “and و” is marked as “not-translated” and “correct”.

و- حسب- نا الله و- نعم الوكيل
Allah is sufficient and He is the best protector

5.9.2 Conjunctions without “and”

Conjunctions other than “and” are illustrated in the following example:

Example of “neither… nor”, “either…or”:

يعني ان- -ك تعمل اشياء انت بتحب تعمل- -ها و- -مش هي مفروضة عليك- -و- -مطلوبة من- -ك او بتعمل- -ها على شان ترضي الناس في- -ها

I mean that you do thing that you love and that are neither imposed on you nor required, or that you do them to satisfy people.

5.10 Prepositions

For prepositions, if there exist equivalent constituents, simply link the correspondent parts.

كام واحد في مصر عند- -هم ثري دي ب- -الاضافة الى الارقام الفلكية دي

In addition to these astronomical numbers, how many people in Egypt have 3D technology?

For cases of two or more propositions in English but only one in Arabic, glue the two prepositions and align them to the preposition counterpart in the other side. For cases of one preposition on one side while no counterpart on the other, the extra proposition will be glued to the object of this preposition even if the object is far away.

Example to glue extra preposition:

ازاي ما- -نحاسب- -ش اعضاء الحزب الغير وطني و- -اللي اتسببوا في خراب البلد ؟

How could we not charge members of* the non National Party who caused the destruction of the country?

For cases where the extra preposition appears after a verb, it is glued to its object despite an obvious co-occurrence of the verb and the preposition. This same rule holds for a preposition following a noun.

Example to glue extra preposition after a verb:

دي كانت حرب و- -هما اللي خلوا- -ها حرب

This was a war and they made it into* a war.

اسمعوا اللي ح- -اقول- -ه كويس

Listen carefully to* what I will say

When the object of an extra preposition is a present participle (verb + ing), align the preposition to the participle.

An example to glue extra preposition to the participle:

يعني في اطار تبادل الاراء ل- -وصول كل من- -نا ل- -شيء يقنع- -ه-

I mean in the framework of* exchanging opinions for each of us to arrive at something that convinces him

5.11 Phrasal Verbs

Phrasal verbs are verbal complexes comprised of a verb and a particle. The particle may be a preposition or an adverb. As the particle-verb combination crucially alters the semantics of the verb, we treat them as a unit for this task (i.e, align the particle with the verb).

Example of verb particle alignment:

طبعاً المقال اللي حضرة- -ك ناقل- -ه لا جدال على- -ه رائع جداً
و- بيتكلم عن الواقع ب- -اسلوب ساخر
Of course the article you copied is excellent without a doubt and it talks about reality in a satirical style.

5.12 Possessives

English possessives can take three forms: "**of**", "**'s**" and " **'** ". The Arabic nominal complementation structure is considered the equivalent of the English possessive construction: "of", "'s" and " ' ". If possessives appear on both sides, simply link them.

Examples to show the alignment of "'s":

ما- -تنسوا- -ش فضل الجيش على- -نا بعد الله
Do not forget the army 's* favor to us , next to Allah.

زي فيلم البرئ ل- -احمد زكي
Like Ahmed Zaki 's movie " The Innocent "

An example to show the alignment of "of":

ازاي ما- -نحاسب- -ش اعضاء الحزب الغير وطني و- -اللي اتسببوا في خراب البلد ؟
How could we not charge members of* the non National Party who caused the destruction of the country?

The only exception to the above rule is when the Arabic word is an adjective implying possession. In this case, the "of" is not tagged with "unmatched and glued"

المصرية المملكة
The Kingdom of Egypt

In Egyptian Arabic the expletive "بتاع/ بتاعة" can function as a literal and direct translation of "of"; thus, they should be aligned

اهي خطبة زي بتاعة اوباما ممكن ابقى مكان- -ه يومين بس
This is a speech like the one of Obama , I may stay in his place for two days only .

English plural possessive consisting of only an apostrophe are treated the same as an apostrophe plus "s" sequence because they are semantically equivalent.

An example to show the alignment of " ' ":

و- -ل- -كم اثارت- -ني الفرحة الغامرة التي اصابت اهل الضحايا ب- -الحكم العبيط
I was amazed at the overwhelming joy of the victims '* relatives after hearing that silly judgment

For extra 's where no equivalent can be found in the source, they are marked as "not-translated correct".

This example shows an instance in which the "'s" serves as a modifier instead of indicating possession.

إنتقدت " هآرتز " أمس في إفتتاحية +ها القرار الذي إتخذت +ه اللجنة المركزية في ليكود
The opening page of yesterday's Haaretz criticized the decision taken by the Likud central committee .

5.13 Passive Sentences

If passive voice appears in both source and target languages, it might be easy to link each piece correspondingly. Whereas for the case of passive voice in the source and the active voice in the translation or vice versa, the word order would be very different and it is not easy to tear apart the corresponding parts. Great care is needed in correctly mapping the links.

Examples to show how to align different cases of passive voice sentences:

و- -الشعب المصري يشغل ب- -موضوع ي- -منتهى التفاهة
And the Egyptian people are engaged by an extremely trivial topic

هل هو كداب فعلاً و- -لا ضحية من ضحايا الثورة اللي اتهاجمت ب- شدة
Is he really a liar or one of the revolution 's victims who have been strongly criticized

5.14 Negation

Negation in Arabic is expressed by different forms using different particles including: ما، لا، لم، لن، ليس. In addition, Egyptian Arabic expresses negation by the structure "ما+فعل+ش". There are often direct translations of the Arabic negation into English.

المشكلة ليست في الاختلاف لان كل واحد عارف هو عايز ايه
The problem is not with differing because everyone knows what they want

يعني مين قتل الجندي مش مشكلة انما المشكلة مشكلة مبدأ
I mean, who killed the soldier is not a problem, but the problem is a matter of principle.

و- -التطاول ان لم يقتل لابد ان يكبر
And if insult is not killed , it must get bigger.

اكيد كل ده ما- -يستاهل- -ش ان القيادة السياسية تتدخل
For sure, all of this is not worth the political leadership intervening

الحل اكيد ما- -يكون- -ش في الاستقالة
Surely, the solution is not resignation

و- -لا اليوم الجيش على ده فعلاً
And I actually do not blame the army for this.

هم- -نفس- -ب حاجة -ش -بيعملوا- -ما-
They do not do anything themselves

اذا مش سافرت و- -كنت في مصر يوم 19 ف- -اكيد ه- -اروح الاستفتاء
If I have not travelled and was in Egypt on the day of the 19th , I will certainly go to the referendum.

بس مش لاقي الرد اللي انا عاوز- -ه
But I can not find the answer that I want

لا يمكن يكون السبب المقنع ل- -شهرة- -ه المفاجئة معارضة- -ه ل- -النظام السابق
The plausible reason behind his sudden fame can not be his opposition to the previous regime

احنا مش ه- -نشارك ف الثورة
We will not take part in the revolution

NOTE: لن <> **will not**

5.15 Relative Nouns

Relative nouns in Arabic relate a subordinate word or clause to the rest of the sentence:(**اللي، الذي، التي، الذين، ما**). In English there are five relative pronouns: that, which, who, whom, and whose. For Arabic relative pronouns, we align them to their English counterparts.

a) Relative noun <> Relative pronoun

الشعب المصري يترك الاهم **الذي** سال من اجل- **-ه** دماء الاطهار
The Egyptian people leave the more important issues for which the blood of the virtuous had been shed

بلاش الافتكاسات **اللي** ما- -ل- **-ها-** -ش طعم
Enough of these goofy theories which lack flavor

ده **اللي** فهمت- **-ه** من توقيع- -ك
That is what I understood from your signature

b) Pronoun <> Relative pronoun

جمعة الغضب الثانية و- **-هي** جمعة غد
The Second Friday of Wrath which is* tomorrow, Friday

مجرد تحليل شخصي اتحمل مسئولية- **-ه***
Just a personal analysis for* which I take responsibility

c) Relative noun <> (none) and vice-versa

Sometimes those pronouns appear in the Arabic source and not in the English translation and vice versa. There are two cases:

Case 1: The relative noun or pronoun refers only to one word. We glue them to their antecedent when it is only one word, and it would be unmatched and labeled with the “glue” tag.

عمر- -ك اشتغلت في المحجر و- -لا شلت **حجارة** و- -علمت في ظهر- -ك
Have you ever worked in a quarry and carried stones which* left marks on your back

Case 2: The relative pronoun refers to more than one word. We treat the relative pronouns as “not translated and correct”, applying the minimum-match strategy.

و- -انت تقارن مصر ب- -باكستان و- -ايران **اللي** لدى- -هما اقوى سلاح في العالم
You should compare Egypt to Pakistan and Iran, having the strongest weapon in the world

5.16 Subordinate Clauses

When the clauses are placed in a hierarchical structure, the more important ones are main clauses and the less important are called subordinate clauses since they cannot stand alone as sentences because they do not provide a complete thought.

A subordinate clause in English is introduced by a subordinate conjunction or a relative pronoun and contains both a subject and a verb. Subordinate conjunctions include *after, although, as, because, before, even if, even though, if, in order that, once, provided that, rather than, since, so that, than, that, though, unless, until, when, whenever, where, whereas, wherever, whether, while, and why.*

Depending on the environment, subordinate clauses can function as subjects, objects, complements and adverbials.

a) clause as subject
   The extra clause markers are treated as "not-translated correct":

المهم اكون انا عارفة انا عايزة اعمل ايه
What is important is to know what I want to do

b) clause as object:

معقول الاعلام المصري قدر يأثر على عقول الناس ب- -الشكل ده
Is it reasonable that the Egyptian media was able to influence the minds of people as such

5.17 Punctuation

All equivalent punctuations can be aligned, including comma-to-comma, period-to-period, and so on. Extra punctuations can be marked as "not-translated" and "correct". Special cases in Arabic are:

Case A: : <> ,

:قال- و- <> **He said ,**

Case B: , <> . (comma on one sentence, period on the other). For this case, link comma to period:

... حيث أضاف ، ه- -استمر في كلام- و-
And he continued talking . He then added ….

Case C: "and" <> "," (or vice versa)

see Section 5.9.1 Conjunctions.

Case D: (none) <> various punctuations (or vice versa)

قبل إنعقاد الإجتماع كان ل- نا لقاء مع الوزير
Before the meeting was held , we met with the minister .

For cases where there is a punctuation mark in one sentence while it disappears in the other sentence, the extra punctuation mark is treated as “not-translated” and “correct”.

6 Special Features in Arabic

6.1 Interrogative Sentence

6.1.1 Usage of “Hal”

The Arabic interrogative words هل have variable equivalents in English language. The question is most of the time asked in English by varying the word-order of the subject and predicate, or by using function words:

هل دولم غير دولكم ؟
Are those ones other than them ones ?

هل تعرف مين المقصود ب- -الباؤن الاحرار ؟
Do you know who is meant by the Freemasons ?

هل الشعب ح- -يجتمع على- -ه ؟
Will he get the people 's consensus ?

هل شفت القطارات في عصر مبارك و- -الحوادث ؟
Have you seen the trains and accidents in Mubarak 's era ?

هل يكون البرادعي ب- -الفعل هو منقذ مصر من توحش ذوي السلطة ؟
Would ElBaradei really be Egypt 's savior from the ferocity of the authorities ?

هل ممكن ان اتمنى ان تسامحوا الضباط ؟
Can I hope that you forgive the officers ?

6.1.2 Interrogative Negation

In cases where you have both interrogation and negation at the same time, we link those which have direct translation in English.

هل ما- سمعتوا- ش- الخبر ؟
Did you not hear the news ?

6.2 VSO versus SVO Structure (subject-drop)

The constructs of “verb before subject” (VSO) and “subject before verb” (SVO) are common structures in Arabic even though classical Arabic tends to prefer VSO. SVO is more common in spoken Arabic, or rather in colloquial Arabic and in less formal MSA (Modern Standard Arabic), and case endings are dropped in pausal forms.

Examples of SVO and VSO alignment:

الطفل أكل التفاحة
أكل الطفل التفاحة
The child ate the apple.

Subject pronouns are normally omitted in Arabic because the subject is marked on the verb.

Example of alignment in the case of a null subject:

مش ه- اسيب مبارك
I* will not insult Mubarak

Note, however, that the subject may be used for emphasis or when using a participle as a verb where the participles are not marked for person. In this case the subject pronouns are aligned to each other.

انا شايف ما- في- ه- ش مشكلة
I think there is no problem

7 Special Features in Egyptian Arabic

7.1 Interrogative Sentence for Yes/No Questions without “Hal”

In Egyptian Arabic, it is more common to ask Yes/No question without using “Hal”. These questions are aligned as if they were affirmative sentences with the English equivalent.

انت شايف جمال مبارك حلو ؟
You think that Gamal Mubarak is good?

هو مصنوع من قماش و- -لا حاجة تانية ؟
Was it made of material or something else?

فاكرين سميرة موسى ؟
Do you remember Samira Moussa ?

بدأتوا تفهموا و- -لا لسة ؟
Have you started to understand or not yet?

7.2 Usage of "في- -ه"

In Egyptian Arabic the structure "في- -ه," sometimes incorrectly transcribed as just "في," should be directly aligned with "There is" without gluing. Its negative equivalent "ما- -في- -ه- -ش" is also treated the same.

حاجات كتيرة في- -ه
حاجات كتيرة في
There are a lot of things.

استقرار ما- -ش
There is no stability

The only exception to this rule is when "في- -ه" has a direct English match

نا- -ه بلاوي في جامعات-
They have issues in our universities

7.3 Typos versus Spelling Variants (due to Egyptian features or genre features)

Due to the absence of a writing standard in Egyptian Arabic, spelling variants are expected. However, these are different than typos. The following rules should be applied depending on the cases to distinguish typos from spelling variants:

Case 1:

A regular typo is any instance when a letter is either added or removed, or a letter is replaced by another. These usually arise from typing mistakes. These should be tagged with "typo"

ب- -سرعة لم تشد- -ها دولة اخرى
With speed that no other country will have witnessed

This includes cases when the letter "ه" at the end of a word was converted to "ة" as this conversion takes away a pronoun:

اللي معاة قرش محيرة يجيب حمام ويطيرة
He who has a piaster that is puzzling him, can buy pigeons and let them fly

On the other hand, when "ة" is converted to "ه", it is considered a spelling variant and is not marked as a "typo"

اعادة الثوره غداً في- -ها شيء مقلق
Repeating the revolution tomorrow has something worrying about it

Case 2:

Instances when extra vowel letters (ا، و، ي، ة) are added to words in a way that does not change how the word is pronounced should NOT be tagged with "typo"

انتة على الهوى يا- -بابا
You are* on the air Dad

تقولو في اول الثورة احنا مش خ- -نشارك ف الثورة
You* say at the beginning of the revolution, we will not take part in the revolution

Case 3:

Anytime one of the following letter pairs was substituted " ت/ط، د/ض، ذ/ظ، ز/ظ، س/ص، ق/ك" the resulting word is considered a typo

عندما كنت اتالع المواضيع الخاصة ب- -جماعة الاخوان المسلمين
As I was looking through topics about the Muslim Brotherhood

Case 4:

Instances when the letter "ق" is substituted with "ا" is deemed to be a spelling variant; therefore, should NOT be counted as a typo.

سألت السؤال ده ابل كده
I have asked this question before

يروحوا يأطعوا- -هم
They will go tear them

Case 5:

Whenever a letter is repeated for emphasis, the word is considered to be a spelling variant; accordingly, it should NOT be tagged with “typo”

لاااااااااااا قـالـوا
They said nooooooooo
They said no

7.4 Alternate Translation

Egyptian Arabic is full of its very own expressions, which sometimes cannot be translated literally. Accordingly, the English translation, in these cases, will present two translations: one with the literal meaning, and the other with the intended meaning. The Arabic expression should be aligned with its English equivalent following the minimum-match approach (word-for-word) if possible. All words in the intended meaning translation should be marked as “not translated and correct” if the meaning of the expression is understood. It should be marked as “not translated and incorrect” if a wrong expression is used. Regardless of the correctness, all words within this translation should be tagged with “Alternate translation.” The square brackets and the straight slash used within this translation should be tagged as “Meta Word.”

Case 1: A correct meaning present:

انـت مـا- -ل- -ك و- -مـال ام شنب- -ه
What do you care about his [MET bloody ALT moustache ALT |MET moustache 's* mother]MET

Case 2: An incorrect meaning present:

حبيب قـلـب- -ه مـيـت و- -ايـد- طرشة -ه
Habib has a dead heart and [MET an ALT iron ALT fist ALT |MET a dumb hand]MET

7.5 Tokenization Errors

The following items should be tagged with a tokenization error whenever present:

Case 1: Null Tokens

Instances when a random token appears that does not correspond to anything in the original file should be marked as “not translated and correct” and tagged with “Tokenization Error”

امـا انـا ف- -لـدى- -ي مـةتُـيـنـثْ سؤال ل- -اي عضو مـن اعضاء هذه الـجمـاعة
As for me, I have a question to any member of this group

Case 2:

Whenever a word normally ending with "ة" gets added to another word, it becomes converted to "ت". When such combined word is tokenizes, the "ت" should revert back to "ة"; otherwise, this is considered a tokenization error.

A correct example:

يا ريت نفك- -نا من حوار ثروة مبارك و- -محاكمة- -ه

I wish we let go of the debate around Mubarak 's fortune and his trial

An Incorrect Example:

اتحمل **مسؤوليت-** -ه

For which I take responsibility

Case 3:

A lot of words in Egyptian Arabic consist of a group of separate parts (tokens), each by itself has a coherent understandable unit of meaning. If these words are tokenized correctly, then the tokens should NOT be tagged with "tokenization error" and aligned with its English equivalent. Similarly, if the word does not get tokenized at all, it would be aligned to all of its equivalent tokens, and NOT tagged with "tokenization error". On the other hand, if such words get tokenized incorrectly so that each token by itself does not present a coherent unit of meaning, then then only these tokens should be tagged with "tokenization error" and aligned as appropriate. This is mostly prevalent in negation, and preposition+pronoun.

A correct example:

اتكلم ب- -العامية المصرية -ي -ل-* **اسمحوا-**

Permit me to speak in colloquial Egyptian

An Incorrect Example:

نتيجة استفتاء الدستور -ش **-عجبهم-** -ما

They did not like the result of the constitutional referendum

Case 4:

In Egyptian Arabic the letter "ب" gets added before verbs to give a sense of immediacy that is implied and cannot be translated into English, the "ب" is considered part of the verb, and should not be present in a separate token. All instances in which the "ب" was separated from its verb, both token should be tagged with "tokenization error"

ليه مش **ب- -تقول** كمان اعداء الوطن ؟

Why do not you* also say enemies of the nation?

This is different from the letters “ح and ه” which serve as an indication of the future, and should be aligned with “will” or “would” if present in English.

لـكن طبـعـاً مـش ه- -يـعـدوا- -هـا كـدا
But of course they will not let it pass like that