Non-compositionality of Idioms in a Multilingual Context [WG1, WG3]

Katalin Ilona Simkó[1], Veronika Vincze[1,2]
[1]University of Szeged
Department of Informatics
[2]MTA-SZTE Research Group on Artificial Intelligence
kata.simko@gmail.com
vinczev@inf.u-szeged.hu

The most basic characteristic of multi-word expressions is their non-compositional meaning, which makes them hard or even impossible to translate word-by-word from one language to the other. Despite this, when working with a pair of languages, certain MWEs seem to behave differently to what we would expect, similarly to what Bogantes et al (2015) showed for dialects of Spanish. This poster aims to show a possible grouping of MWEs that would be beneficial for machine translation as well as to offer an overview of a database for English-Hungarian idioms in this framework.

We are distinguishing five different ways an idiom and its translation may be related:

(a) there is an idiom expressing the same meaning in both source and target language with the same basic syntactic and lexical structure

(b) there is an idiom expressing the same meaning in both source and target language but there is a slight change in either the syntactic or the lexical structure

(c) there is an idiom expressing the same meaning in both source and target language, but with different syntactic and lexical structure

(d) the literal translation of an idiom is also an idiom in the target language, but with a different (additional) meaning to what is in the source language

(e) there is no idiom in target language expressing the same meaning

These five types require different treatments in machine translation and they are also very interesting from a linguistic viewpoint. While all idioms have non-compositional meaning within a certain language, some seem to be compositionally or semi-compositionally translatable for certain language pairs while others become nonsensical when translated word-for-word or have a very different meaning.

In our English-Hungarian database, four of the above mentioned five types of idioms are listed with their translation. We collected idioms from the web (http://www.sulinet.hu/nyelvek/?nyelv=angol) and then classified them into the above categories. For type (a), some examples are easy to trace back to a common source, like (1), which comes from the Bible, while others are not so straightforward, like (2).

(1) EN to cast the first stone
HU első követ veti rá
first stone-ACC cast-PRES.3SG him-SUB
'casts the first stone on him'
"be first to attack/blame someone"

(2) EN to break the ice
HU megtöri a jeget
PERF.break-PRES.3SG the ice-ACC
'he breaks the ice'
"he gets the conversation started"

The idioms in type (b) differ either in vocabulary, like (3) or in structure, (4):

(3) EN to kill the goose that layed the golden egg
HU megöli az aranytojást tojó tyúkot
PERF.kill-PRES.3SG the goldenegg-ACC laying hen-ACC
'kills the hen laying the golden egg'
"get rid of something profitable"

(4) EN one's bark is worse than one's bite
HU csak ugat de nem harap
only bark-PRES.3SG but not bite-PRES.3SG
'he only barks, but does not bite'
"he is not as bad as he seems"

Type (c) contains pairs of idioms that express the same meaning with different lexical items:

(5) EN let the grass grow under one's feet
HU megalszik a tej a szájában
PERF.curdle.PRES.3SG the milk.NOM the mouth-POSS3SG-INE
'the milk curdles in his mouth'
"he is very slow"

For type (d) only a few examples are yet discovered, so we can claim that finding the same idiom with a different meaning in two languages seems to be a rare occurrence. For these, both idioms and their meaningful translation to the other language are contained in the database. While the English idiom in (6) expresses that someone got very scared, its literal translation to Hungarian means that someone is really excited.

(6) EN jumps out of one's skin
HU kiugrik a bőréből
out.jump.PRES.3SG the skin-POSS3SG-ELA
'jumps out of his skin'
"he is very excited"

Type (e) contains idioms that are only present in the source language and not expressible idiomatically in the target language. The idiom in (7) is only available in English, while the one in (8) only in Hungarian.

(7) EN penny-wise, but pound-foolish
"careful about small amounts of money, but wasteful with big amounts"

(8) HU nem erőszak a disznótor
no violence-NOM the pig-killing-NOM
'the pig-killing is no violence'
"you are free to refrain from an invitation"

This paper gives a basic introduction to a classification system of multi-word expressions in a multi-language context and presents a database of English-Hungarian idioms that was constructed following this classification system, together with statistical data on the categories. In the future, we intend to expand our research in multiple directions: adding more idioms to the current database, including more types of MWEs and working with more languages, as we hope that this classification could enhance machine translation and other linguistic research.

References

Bogantes, D., Rodríguez, E., Arauco, A., Rodríguez, A., Savary, A.: Towards Lexical Encoding of Multiword Expressions in Spanish Dialects (2015) Poster. PARSEME 5th General Meeting.