Technical Reports on Language Technology
Report No 25, 20018
http://www.njas.helsinki.fi/salama

Normalizing English for Interlingua:
Multi-channel Approach to Global Machine Translation

Abstract

The paper tries to demonstrate that when English is used as interlingua in translating between two languages it can be normalized for reducing unnecessary ambiguity. Current usage of English often omits such critical features as the relative pronoun and the conjunction for marking the beginning of the subordinate clause. In addition to causing ambiguity, the practice also makes it difficult to produce correct structures in target language. If the source language makes such structures explicit, it is possible to carry this information through the whole translation chain into target language. If we consider English language as an interlingua in a multilingual translation environment, we should make the intermediate stage as little ambiguous as possible. There are also other possibilities for reducing ambiguity, such as selection of less ambiguous translation equivalents. Also long noun compounds, which are often ambiguous, can be presented in unambiguous form, when the linguistic knowledge of the source language is included.

1 Abbreviations

This SL = source language
TL = target language
MT = machine translation
RBMT = rule-based machine translation
EBMT = example-based machine translation
SMT = statistical machine translation
NMT = neural machine translation
POS = part-of-speech
MWE = multiword expression
[ENG-ORG] = original English
[SWA] = Swahili language
[INTL] = interlingua
[SA] = translation using Salama

2 Introduction

Two things have motivated me to write this paper. One was the observation that, when translating from Swahili via English into Finnish, the translation result was often better than when translating the same sentence from current English into Finnish. That is, when English was used as interlingua, translation results were better than when translation was done directly from one language into another. The other thing is that I am worried about the fate of the large majority of languages, if we continue to allocate almost all funds to statistical and neural machine translation efforts. These methods would be suitable for not more than about one percent of world's languages. The rest of them, 99 percent, would be left without proper translation technology.

At first glance, these two things do not seem to have anything in common. In this paper I try to show that they do have much in common. For less resourced languages we could use rule-based methods, which currently are neglected due to the high cost of technology development. The findings discussed in this paper encourage to develop translation technology based on the use of interlingua, which is here considered as normalized English.

Translation issues are discussed here in the context of two translation systems included in the Salama Translator (ref. to be added). Both of the translation systems, which were used in translating from Swahili to English and from English to Finnish, are strictly rule-based. Therefore, they make maximal use of detailed linguistic information. Translation result through interlingua can be considerably improved by 'normalizing' the form of English that is used as interlingua. A big part of the normalization process includes the selection of sentence structures and words, which have the minimal amount of ambiguity. Another possibility for improving translation quality is to preserve part

of the linguistic codes of the source language (SL) and compare them with the codes produced by the analyzer of English. These codes can then be disambiguated in the environment, where context-sensitive rules can be written. This procedure is demonstrated in section 6.

Current English, as it appears in news texts, often omits relative pronouns and conjunctions initiating a subordinate clause. These omissions alone are a major source of confusion when translating from English. In target language (TL), these omitted features must be made explicit, however. It is a well-known fact, that it is easier to delete something than to create new out of nothing. This, however, must be done when translating from current English into another language.

The omission of clause boundary markers in English causes constant disambiguation problems even on such elementary level as POS marking. As is commonly known, English is highly ambiguous regarding POS resolution. If POS tagging makes mistakes, it distorts syntactical mapping and all other phases in translation process. Therefore, all measures that could be used for improving the accuracy of POS tagging should be used.

The problems caused by omissions can be avoided, or at least reduced, when all relevant linguistic features from SL are retained in translation process for further use. For example, Swahili encodes relative pronouns in detail, and uses systematically conjunctions, or other corresponding structures, in initiating subordinate clauses. All this information can be transferred to English translation, which makes the translation into the third language easier.

Another method for reducing ambiguity is the choice between two or more alternative words in target language. For example, the words *that* and *which* function as relative pronouns. The word t*hat* is highly ambiguous and could be avoided altogether in the function of relative pronoun. This word could be reserved for initiating subordinate clauses, and for demonstratives. The word *as* has many meanings, and it should be avoided for marking the beginning of temporal subordinate clauses. The less ambiguous conjunction *when* should be used instead.

Ambiguity can be reduced also by selecting such verb structures, which are less ambiguous. Above we had a sentence: *A big part of the normalization process includes **the selection** of such sentence structures, which have the minimal amount of ambiguity*. This sentence could be rephrased as: *A big part of the normalization process includes **selecting** sentence structures, which have the minimal amount of ambiguity*. Here the gerund form *selecting* can justifiably be interpreted as adjective, although it could also be a verb, or noun. If we use infinitive form *(to select*) instead of gerund, we avoid ambiguity. Gerund and infinitive are not equivalent, of course, but we should avoid gerund when its use is not necessary.

One might argue that by 'normalizing' English we do not get very fluent language. English is here only in the intermediate role, and the only thing that matters is that the message of the SL will be translated into the TL as accurately as possible.

It has been suggested that translation quality could be improved by simplifying source text sentences and structures (Hasler et al, 2017). Long sentences are known to be problematic for neural systems. For RBMT such problems do not occur, because the processing methods are different. It would also be very difficult to construct such an automaton for cutting sentences, which would be reliable. When we normalize text as we discuss in this paper, we do not delete any information. We just try to make the text more computer-friendly.

3 Examples of text normalization

Below I will demonstrate the problems discussed above with examples from the news text challenge of the WMT17[1].

```
Trump says he was being sarcastic when
claiming Obama was the founder of Isis.
```

When this is translated into Finnish without attempting to add equivalences for words, which were omitted in the sentence, we get the incorrect translation as in (1).

(1) *Trump sanoo hän oli sarkastinen väittäessään Obama oli Isisin perustaja.*

Now we take the same sentence in Swahili as starting point (2).

(2) `Trump anasema, kwamba alikuwa mwenye kejeli alipodai, kwamba Obama alikuwa mwanzilishi wa Isis.`

This is translated into English as in (3).

[1] www.statmt.org/wmt17/translation-task.html

(3) `Trump says, that he was sarcastic when he claimed, that Obama was the founder of Isis.`

When this is translated into Finnish, the result is correct (4).
(4) *Trump sanoo, että hän oli sarkastinen kun hän väitti, että Obama oli Isisin perustaja.*

The missing relative pronoun may also be in the position of an object or indirect object, as in (5).
(5) `The country however was not mentioned among the states one can call directly from the room.`

The verb *call* is ambiguous in that it means *to invite* and *to make a telephone call*. In this context it obviously means the latter, but it is very hard to disambiguate. We give a Swahili version of the sentence and its translation into English (6).
(6) `Hata hivyo, nchi haikutajwa kati ya nchi zile, ambazo kutoka kwake inawezekana kupigiwa simu moja kwa moja kutoka katika chumba.`
`However, the country was not mentioned among those countries, from which it is possible to call directly from the room.`

When this interlingua version of the sentence is translated into Finnish, we get a reasonable translation (7).
(7) *Maata kuitenkaan ei mainittu niiden maiden joukossa, mihin voi soittaa suoraan huoneesta.*

Below are more examples of similar cases. Abbreviations show the version in question (8).
(8) [ENG-ORG] `According to analysts, it is difficult to predict the extent of the effect the people's rage will have on the election result.`
[SA] *Analyytikoiden mukaan on vaikeaa ennustaa vaikutuksen määrän ihmisten raivolla on vaalitulokseen.*

[SWA] `Kwa mujibu wa wachambuzi ni vigumu kutabiri upana wa athari, ambayo ghadhabu ya watu itakuwa nayo kwa matokeo ya uchaguzi.`
[INTL] `According to the analysts it is difficult to forecast the width of the effect, which the rage of people will have for the results of the election.`
[SA] *Analyytikoiden mukaan on vaikeaa ennustaa sellaisen vaikutuksen määrää, joka ihmisten raivolla on vaalin tuloksiin.*

(9) [ENG-ORG] `There were four bombs that exploded in a tourist attraction also favoured by the Finns, reports Thai newspaper Bangkok Post.`
[SA] *Oli neljä pommia, jotka räjähtivät matkailuvetonaulalla suosivat myös suomalaiset, raportoi thaimaalainen sanomalehti Bangkok Post.*

[SWA] `Kulikuwa na mabomu manne, ambayo yalilipuka katika senta ya watalii, ambayo pia ilipendwa na Wafini, inaripoti Bankok Post.`
[INTL] `There were four bombs, which exploded in the tourist centre, which also was liked by the Finns, reports Bankok Post.`
[SA] *Oli neljä pommia, jotka räjähtivät turistikeskuksessa, mistä myös pitivät suomalaiset, raportoi Bankok Post.*

A particularly problematic case is the sentence below, where the omission of critical words makes the sentence highly ambiguous (10).
(10) `They argued the makeshift shops and restaurants often provided shelter and free meals to those in need.`

This could be understood in either of the following ways.
(11) Either: `They argued the makeshift shops and restaurants, which often provided shelter and free meals to those in need.`
Or: `They argued that the makeshift shops and restaurants often provided shelter and free meals to those in need.`

Probably the latter interpretation is correct, but the sentence structure does not tell it in any way. Now again, when we start from Swahili, we get the following versions.
(12) [SWA] `Walidai kwamba maduka ya muda na mahoteli mara nyingi zilitoa kivuli na chakula bure kwa maskini.`
[INTL] `They claimed that the provisional shops and hotels many times provided shade and free food to the poor.`
[SA] *He väittivät, että väliaikaiset kaupat ja hotellit monta kertaa tarjosivat suojaa ja vapaan ruoan köyhille.*

4 Long noun compounds

It is well known that long noun compounds in English are often ambiguous, because the language does not encode the internal structure

of the noun chain in any way. Consider the example in (13).

(13) [ENG-ORG] `It is possible to prevent forestry capital income losses.`
[SA] `On[V] mahdollista[A] estää[V] metsänhoito[N] pääomavaltaisia[A] tulo--[N] menetyksiä[N].`

[SWA] `Inawezekana kuzuia hasara za mapato ya rasilimali za elimumisitu.`
[INT] `It is possible to prevent the losses of the incomes of the capital of the forestry.`
[SA] *On mahdollista estää metsänhoidon pääoman tulonmenetyksiä.*

Note that the original English text was translated with Salama without using the component, which identifies multiword expressions. The translation with POS codes shows how the system interpreted the sentence. The word *capital* was interpreted as adjective, and *income* was interpreted as being the first member of the compound *income loss*.

When we start the process from Swahili, we get a precise structure of the noun compound. Swahili uses the genitive compound structure in expressing noun compounds. In addition, it has a noun class system, which makes it possible to define the referent of each compound member by encoding the referent in the genitive connector. By default, Salama translator produces the corresponding genitive structure in English. It is not ideal English, but it has all information in the form, which is easy to translate into TL.

In (14) we have a still more complex noun compound. The noun compound with four members can be described as a MWE, and in fact, this is the only solution when we translate from original English. Below we demonstrate, that it is not necessary, if we start translation from Swahili.

(14) [ENG-ORG] `Last week we bought a single family house lot.`
[SA] *Viime[DET] viikolla[N] me[PRON] ostimme[V] yksittäisen[A] perhe--[N] talon[N] paljon[N].*

[SWA] `Juma lililopita tulinunua kiwanja cha nyumba ya familia moja.`
[INTL] `Last week[ADV] we bought[V] the plot[N] of the house[N] of one family[N].`
[SA] *Viime viikolla me ostimme yhden perheen talon tontin.*

We see that the word *lot* is highly ambiguous, and it gets easily mistranslated. When we translate the sentence from Swahili, we get the representation with genitive components. The translation from interlingua into Finnish gives understandable result, although the expression *yhden perheen talon tontin* can be glossed also as a real Finnish compound *omakotitalotontin*.

The aim of modifying English as interlingua is not to make it in any way artificial. In fact, we try to return it back to the stage, where clause boundaries are clearly marked. We also try to avoid the use of such words and wordforms, which repeatedly cause translation problems, due to their high degree of ambiguity.

5 Why English as interlingua?

The ideal interlingua is a system, which describes all linguistic features, such as POS, syntax, and semantics in such a general way, that it can be used as an intermediate step in translating between any two languages, Esperanto has been used for this role, and also fully artificial interlinguas have been suggested (Dorr et al, 2006). In the 1990s it was thought that it could be possible to convert texts into formal representations common to more than one language (Hutchins, 1995). Also for special domains, such as bio-medical information, interlingual frames for mapping between clinical vocabularies have been suggested (Masarie et al, 1991).

The POS tagging is probably the most universal feature in languages, but variations start in syntax, and the lexicon of each language is almost unique. It has been suggested that, instead of lexical representation, semantic description should be used. The problem is, however, that we cannot describe semantic content without using words. And in any case, we must convert that semantic content into words in target language.

Why should we use English as interlingua, although it is not ideal for that purpose? Despite its weaknesses it has many advantages. Many people know the language and there is also a need to construct machine translation systems between English and another language in any case. Some form of English functions as interlingua in many multilingual translation systems already (Virk et al, 2014, Ranta, 2004, 2011), although neural systems might be able to avoid a real and tangible language as interlingua.

6 Using linguistic information from source language

In addition to normalizing English for interlingua, we can also use another method, which is based on the linguistic information inherited from the source language. Because the translation system discussed here is based on full linguistic description, it is possible to retain in translation any type of linguistic information, which is considered useful in translating into a third language.

Also, for example-based MT has been suggested a mechanism, where the meaning content of the source text is preserved throughout the translation process (Chong et al, 2017). Yet it is complex and challenging, because EBMT does not encode language in detail. For RBMT this is considerably easier because of its covering and accurate encoding.

```
Raid alisema Alhamisi, kwamba anatabiri,
kwamba Clinton atashinda katika
uchaguzi.
```

The interlingua representation of the Swahili sentence can be presented in various ways. In (15) below, POS tags are added in translation after a single word or after a multiword expression.

(15)
```
Raid said [V] Thursday [N], that
[CONJ] he forecasts [V], that [CONJ]
Clinton will win [V] in [PREP] the
election [N].
```

Perhaps a better representation, which would make the analysis of the interlingua (i.e. English) easier, would be to join the word and tag together, such as in (16).

(16)
```
Raid said[V] Thursday[N], that[CONJ]
he forecasts[V], that[CONJ] Clinton will
win[V] in[PREP] the election[N].
```

It is also possible to add syntactic tags to translation, as in (17).

(17)
```
Raid[@SUBJ] said[@FMAINVtr+OBJ>]
Thursday[@SUBJ], that he
forecasts[@FMAINVtr-OBJ>], that
Clinton[@SUBJ] will win[@FMAINVtr-OBJ>]
in[@ADVL] the election[@<P].
```

Note that not every word has a syntactic tag. This is due to the word structure of Swahili, where finite verbs encode also such features as subject pronoun, relative pronoun, and object pronoun. In such cases there is only the syntactic tag of the verb itself. Also the auxiliary verb 'will' is without tag, because the future tense in Swahili is encoded into the main verb as a prefix (*a-**ta**-shinda*).

It is also possible to include both POS tags and syntactic tags to the translation, as in (18).

(18)
```
Raid[@SUBJ] said[V][@FMAINVtr+OBJ>]
Thursday[N][@SUBJ], that[CONJ] he
forecasts[V][@FMAINVtr-OBJ>], that[CONJ]
Clinton[@SUBJ] will win[V][@FMAINVtr-
OBJ>] in[PREP][@ADVL] the
election[N][@<P].
```

Tests show that keeping a large number of tags in the translation process makes processing slower. It may be that only part of the tags is needed in further processing. The selection of tags to be included can be done easily, and experience will show, which tags are really needed.

Below we will see an example, where the retention of some tags of source text are useful.

(19) [ENG-ORG]
```
Ingredients for a
disaster - see the picture gallery of
the Parafest destruction.
```
[SA]
```
Ainekset[N] katastrofiin[N] -
näkevät[V][PRES] kuva[N] Parafestin[N]
hävityksen[N] gallerian[N].
```

[SWA]
```
Viambato kwa maangamizi - tazama
picha za uangamizi za Parafest.
```
[INTL]
```
The Components[N] for[PREP]
destruction[N] - look[V][IMP] the
pictures[N] of destruction[N] of
Parafest[PROPNAME].
```
[SA] *Ainekset[N] katastrofiin[N] - katso[V][IMP] kuvagalleriaa[N] Parafestin[N] hävityksestä[N].*

It may be argued that it will not be possible to handle the mixture of words and codes in an analyzer, which is designed to analyze normal text. This is true. But is it possible to modify the analyzer so that it performs the analysis ignoring the codes when analyzing, but retains the codes throughout the process? The answer is, yes. At least in finite state analysis systems this is simple, provided that we use the representation, where the word and codes attached to it are joined as a single string, as in (19) above. Finite state systems are normally designed so that they handle strings separated by the word boundary code (i.e. space, or line boundary on left or right). The finite-state lexicon can be constructed so, that the word has, in addition to terminating at the end of the word proper, also continuation to such words, which have linguistic codes added to them. The analysis goes on without being

affected by the codes, and the analyzer of English adds new codes, as the system requires.

Now it turns out that after analysis and disambiguation there are two separate codes for the same phenomenon, one inherited from the source text and another inserted by the analyzer of English. If the codes are identical or otherwise identifiable as encoding the same phenomenon, the analysis can be accepted as correct. In case the codes are different, we must choose the correct one. Current English is notoriously difficult to disambiguate even on the basic POS level, especially when important relative pronouns and subordinate clause markers (including commas) are omitted. Here the codes inherited from source language will help tremendously. For example in Swahili, verbs, nouns and adjectives are separate categories, identifiable by morphology and location in sentence.

When we get a representation of the sentence with codes inherited from the source language, we can construct a disambiguation system, where, when writing disambiguation rules, we can take the context in consideration. Consider the example in (20).

(20) `Originally many used spears for protecting themselves.`

When this is analyzed with the current analyzer of English, the result is as in (21)

(21)
```
Originally[ADV][@ADVL]
many[DET][@A>] used[A][@A>]
spears[N][@NH] for[PREP][@ADVL]
protecting[V][@-VFIN-ing]
themselves[REFL][@OBJ]
```

When the word *used* is interpreted as adjective, the translation fails badly. If we take the Swahili version (22) of the same sentence, we get the English translation as in (23).

(22) `Awali wengi walitumia mikuki kwa kujilinda.`

(23)
```
Originally[ADV][@ADVL]
many[PRON][@SUBJ]
used[V][@FMAINVtr+OBJ>] spears[N][@OBJ]
for[PREP][@ADVL] protecting
themselves[N][@<P].
```

On the basis of this, we get the correct translation in Finnish (24).

(24) *Alkujaan monet käyttivät keihäitä suojelemaan itseään.*

7 A word on global machine translation

Current multilingual machine translation systems are usually based on a single architecture. Apart from some promising academic test systems, they are owned by large global companies, such as Google and Microsoft. For them, machine translation is a side track, and the current status shows that only big languages are included into the translation system. Furthermore, they make use of large masses of human-translated parallel texts, which can be manipulated and converted into translation systems. Statistical systems, and more recently neural systems (Cho et al, 2014), make use of these resources. Because these systems rely on increasing computer power, without need to invest in paid expert people, they are very attractive for commercial companies, and also for universities. Therefore, labor-intensive rule-based translation systems do not play a role there. As a result, over 99% of world's languages will be left without machine translation systems. It is not cost-effective to develop them. And more important, it is not even possible to develop translation systems for those languages, because they do not have required masses of parallel texts. The current trend of investing in neural systems speeds up the marginalization of the big majority of languages.

Statistical and neural systems have well-known weaknesses. They include the long sentence problem (Hasler et al, 2017), rare word problem (Luong et al, 2015), and above all, the absence of linguistic knowledge. Various methods have been suggested for including some linguistics into the system, such as hybridization (Labaka etal, 2014, Habash et al, 2009), a language model (Gulcehre et al, 2017, Nielsen and Ney, 2004), context-dependent word representation (Choi et al, 2017), and a fine-grained attention mechanism (Choi et al, 2018). It remains to be seen how efficient these improvements are. It has also been suggested that rule-based and statistical systems could be integrated (Zbib et al, 2012).

In addition to statistical and neural systems, there are some promising rule-based systems with the aim to develop into multilingual unified systems. These include the linguistic development environment called Nooj (Silberztein. 2016), and the grammatical framework approach (Virk et al, 2014, Ranta, 2004, 2011). Both of these systems have already included several languages into the same design architecture.

Although I have developed rule-based translation systems without following any of the other existing models as such, they have much in common with those. Any rule-based approach is likely to address the same problems, and solutions must be found using similar criteria.

My approach is different in that it does not require the translation system to follow a certain model. We could think about the global translation system as a network of languages, where each language has a high quality translation system between the local language and English, to both directions. The translation systems can be of any brand. The only thing that matters is quality. There can be rule-based, example-based, phrase-based, or neural systems in the palette.

It would be advisable that the translation into English could produce normalized translation, as discussed above. With rule-based systems it is quite easy, because the output can be controlled. For statistical and neural systems it might be more difficult.

If we allow all kinds of translation systems to be part of the global system, the requirements for the form of English must be considered from the viewpoint of various types of translation systems. For example, the inclusion of linguistic codes from source language may be relevant only in rule-based systems, because their encoding system is reliable. To require linguistic encoding from statistical or neural systems might be unrealistic. Even the normalization process, as discussed above, might be impossible for those systems.

As a result, we can foresee that one percent of world's languages will use statistical and neural methods, and 99 percent of them will rely on rule-based methods. It may be even possible that some of the one percent will use rule-based methods. Swahili and Finnish, discussed here, belong to that one percent, but rule-based systems here are highly competitive with statistical methods, and outperform them in many respects. This can be explained by the fact that both Swahili and Finnish are morphologically very complex, and it is in practice impossible to handle all word-forms without analyzing each wordform.

8 Conclusion

In this paper I have argued that a normalized version of English could be used as interlingua in multilingual translation environment. Part of the ambiguities inherent in current English text can be avoided, if the translation from SL into English is geared to produce such translation, which avoids ambiguity as far as possible. The inclusion of markers for relative clauses and subordinate clauses is one example. Also the selection of translation equivalents helps in disambiguation. Long English compounds can be made more transparent, when information on the compound structure is carried from SL into the interlingua. Selection and inclusion of codes from SL can also help in disambiguation of the English text.

These issues were illustrated with examples from the WMT17 translation challenge of news texts.

The paper does not claim that translation through interlingua would avoid problems inherent when we translate from current English. It only shows that having an intermediate language between two languages does not always multiply translation problems. In some cases it makes it easier. And these cases are often occurring phenomena, which make rule writing a nightmare.

It was also suggested that by concentrating on developing high quality translation systems between local language and English we could achieve a global translation system, where all types of translation systems could contribute.

References

Cho Kyunghyun, van Merrienboer Bart, Bahdanau Dzmitry and Bengio Yoshua. 2014. On the Properties of Neural Machine Translation: Encoder--Decoder Approaches, Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.

Choi Heeyoul, Kyunghyun Cho, Yoshua Bengio. 2017. Context-dependent word representation for neural machine translation. Computer Speech & Language, 45:149-160.

Choi Heeyoul, Kyunghyun Cho, Yoshua Bengio. 2018. Fine-grained attention mechanism for neural machine translation. Neurocomputing, 284:171-176.

Chong Chai Chua, Tek Yong Lim, Lay-Ki Soon, Enya Kong Tang, Bali Ranaivo-Malancon. 2017. Meaning preservation in Example-based Machine Translation with structural semantics. Expert Systems with Applications, 78:242-258.

Dorr B., E. Hovy, L. Levin. 2006. Machine Translation: Interlingual Methods. Encyclopedia of Language & Linguistics (Second Edition), Pages 383-394.

Gulcehre Caglar, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Yoshua Bengio. 2017. On integrating a language model into neural machine translation. Computer Speech & Language, 45:137-148.

Habash N., Dorr B. and Monz C. 2009. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. Machine Translation, 23(1):23-63.

Hasler Eva, Adrià de Gispert, Felix Stahlberg, Aurelien Waite, Bill Byrne. 2017. Source sentence simplification for statistical machine translation. Computer Speech & Language, 45:221-235.

Hutchins W. John. 1995. Machine Translation: A Brief History. Concise History of the Language Sciences, 431-445.

Labaka G, España-Bonet C., Mà rquez L. and Sarasola K. 2014. A hybrid machine translation architecture guided by syntax. Machine Translation, 28(2):99-125.

Luong Minh-Thang, Ilya Sutskever, Quoc V. Le, Oriol Vinyals and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation. arXiv:1410.8206 [cs.CL].

Masarie Fred E., Randolph A. Miller, Omar Bouhaddou, Nunzia B. Giuse, Homer R. Warner. 1991. An interlingua for electronic interchange of medical information: Using frames to map between clinical vocabularies. Computers and Biomedical Research, 24(4):379-400.

Nielssen S. and Ney H. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. Computational Linguistics, 30(2):181-204.

Ranta, Aarne. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism. The Journal of Functional Programming, 14(2):145–189.

Ranta, Aarne. (2011). Grammatical Framework: Programming with Multilingual Grammars. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Silberztein, Max. 2016. Formalizing Natural Languages: the Nooj approach. Wiley eds.

Virk, Shafqat Mumtaz, K.V.S. Prasad, Aarne Ranta and Krasimir Angelov. 2014. Developing an interlingual translation lexicon using WordNets and Grammatical Framework. Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pp. 55–64.

Zbib R., Kayser M., Matsoukas S., Makhoul J., Nader H., Soliman H. and Safadi R. 2012. Methods for integrating rule-based and statistical systems for Arabic to English machine translation. Machine Translation, 26(1-2):67-83.