ON MACHINE TRANSLATION

CĂTĂLIN DEHELEAN
Babeş-Bolyai University, Cluj-Napoca, Romania

Keywords: *machine translation, natural language*

Abstract

Any machine translation should be user-friendly. Thusly, the perspective of the user ought to be taken into account. Accordingly, any developer should be well-informed about the make up human or natural language. This article is meant to introduce any interested party into the basics of the language levels and their categories which are relevant for a successful machine translation.

Introduction

What does anyone want from a translation engine? The answer is simple: to get a useful translation of the original text. (See Figure 1.)

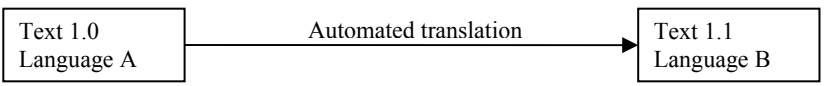

Figure 1: A graphic representation of the public expectation of the automated translation process.

Logic seems to dictate that this would be the point where this article should actually end, and yet it doesn't. The reason for the continuation is rather simple: many a time machine translation does not yield the expected results, as the text it produces is either faulty, in that the language of the translation is broken (See Figure 2)

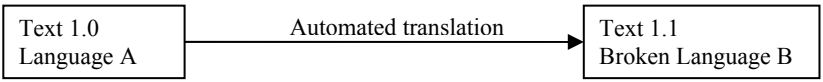

Figure 2: A graphic representation of a type of unsuccessful translation

and/or it contains sequences of the language the original text was written in, (See Figure 3)

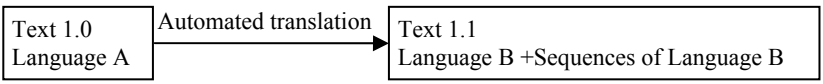

Figure 3: A graphic representation of a type of unsuccessful translation

or utterly unsuccessful, that is the language of the translation is identical to the language of the original text. (See Figure 4)

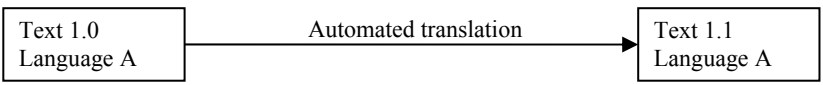

Figure 4: A graphic representation of a type of unsuccessful translation

It may be argued, of course, that the reasons for an unsuccessful machine translation are indeed multiple and complex. Some of the most common and obvious reasons are: technical problems, software bugs and language complexity. (See Figure 5.)

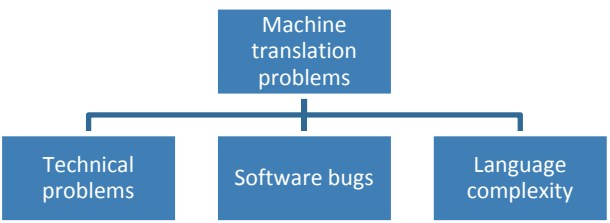

Figure 5: A graphic representation of possible reasons for unsuccessful machine translation

However, it is the position of this article that most if not all of these technical reason are, as a matter of fact, incidental and ultimately not the cause of statistically relevant failure. Nor is the reason connected with the coding. The codes in themselves work just fine. The problem which inevitably leads to repeated and thus consistent failure of machine translation is a poor understanding of the complexity of human or natural language.

Assessment criteria

It all seems terribly complicated, and, at times, it certainly can be tremendous, even if all eyes are pried exclusively on the structure of the human language. But if one is to look at this problem from the average user's angle, the perspective might change. Any natural language can be studied and understood on different levels and, of course, each level presents its own assessment criteria.

Firstly, any language in its natural form is spoken. As such, one is bound to relate to the phonological level. The phonological level has its own phonological criteria. (See Figure 6.)

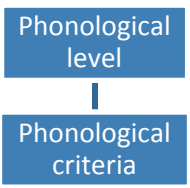

Figure 6: A graphic representation of the relationship between the phonological level and the phonological criteria

Secondly, no language is imaginable without words. Accordingly one ought to take into account the lexicological level. The lexicological level has its own lexicological criteria. (See Figure 7.)

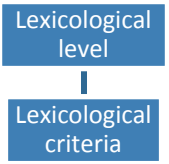

Figure 7: A graphic representation of the relationship between the lexicological level and the lexicological criteria

Thirdly, a language uses various beginnings or endings to modify words. To understand this phenomenon, the study of the morphological level is required. The morphological level has its own morphological criteria. (See Figure 8.)

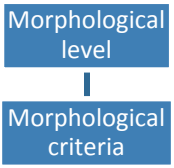

Figure 8: A graphic representation of the relationship between the morphological level and the morphological criteria

Fourthly, the words, in the intended form must be ordered in a certain way to produce an utterance. In this case, one studies the syntactic level. The syntactic level has its own morphological criteria. (See Figure 9.)

Figure 9: A graphic representation of the relationship between the syntactic level and the syntactic criteria

Fifthly, depending on the intention of the speaker, any utterance can have several meanings. Consequently, one can speak about the

semantic level. The semantic level has its own semantic criteria. (See Figure 10.)

Figure 10: A graphic representation of the relationship between the semantic level and the semantic criteria

So there are five language levels and, thus, some aspects regarding the five assessment criteria are to be discussed in this article. (See Figure 11.)

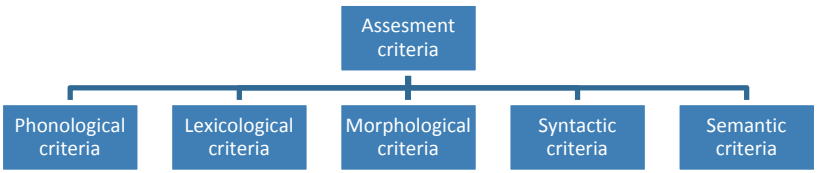

Figure 11: A graphic representation of the assessment criteria for any natural language in automated translation.

Phonological criteria

Machine translation seems not to need them. After all, all the user has to do is to type in a text click 'translate' and the text in the target language is displayed. But every translation engine has a component which enables the user to have the text in either the source or target language be automatically spelled out loud. Admittedly, this component is indeed of secondary importance. However since it is to be found there, it can be the subject of an informed discussion on the linguistic principles which the algorithms ought to respect in order to yield results which any user may deem acceptable.

The first important observation to be made here is that many languages have not one but several standard varieties. (See Figure 12.)

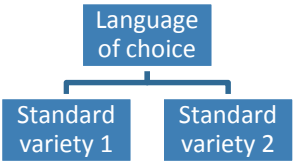

Figure 12: A graphic representation of a language of choice with two standard varieties

As such, the first and foremost thing to do is to choose a standard variety of the language under consideration. (See Figure 13.)

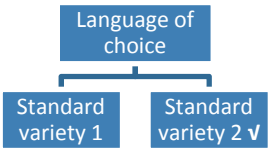

Figure 13: A graphic representation of the choice of a standard variety of a language of choice with two standard varieties

Once the choice of a standard variety of a language is made, there are several criteria one is bound to respect in order to achieve not only a degree of intelligibility, but also broad language functionality. (See Figure 14.)

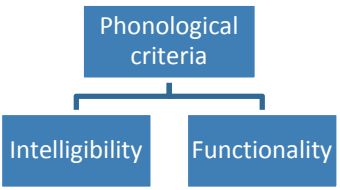

Figure 14: A graphic representation of the purpose of the phonological criteria

While there is a number of phonological criteria to be discussed later in this work, all of the are based on two ideas, namely proper pronunciation and proper intonation. (See Figure 15.)

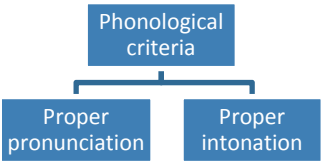

Figure 15: A graphic representation of the relationship between Phonological criteria, Proper pronunciation and Proper intonation

Proper pronunciation is a generic term which encompasses proper pronunciation of the phonemes, proper pronunciation of consonant groups, proper pronunciation of syllables, proper pronunciation of the stress, proper pronunciation of the strong and weak forms and proper pronunciation of contractions. (See Figure 16.)

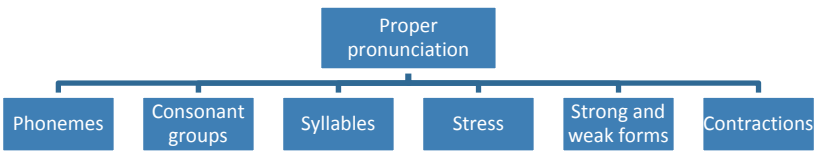

Figure 16: A graphic representation of the proper pronunciation criteria

A further explanation is necessary with regards to proper pronunciation of the stress. It too is generic and can be divided into proper pronunciation of the word stress, proper pronunciation of the phrase stress, proper pronunciation of the sentence stress and pronunciation of emphatic stress. (See Figure 17.)

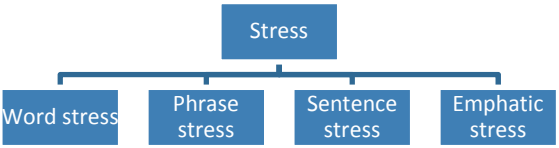

Figure 17: A graphic representation of the divisions of the proper pronunciation of the stress

Lexicological criteria

Lexicological criteria are by and large the most prevalent. Since a text exists only through words, a comprehensive list of words is an imperative. Opinions however may differ on how this list should come to be, but ultimately there are two basic orientations, traditional and modern which are actually complementary and are employed at the same time. (See Figure 18.)

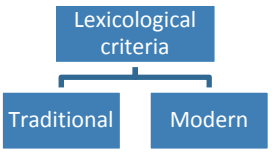

Figure 18: A graphic representation of the types of orientation, typical of lexicological criteria

Firstly, there is the traditional orientation of compiling words into an electronic dictionary. (See Figure 19.)

Figure 19: A graphic representation of the traditional orientation, typical of lexicological criteria

As logic dictates, the electronic dictionary is basically a database of entries. (See Figure 20.)

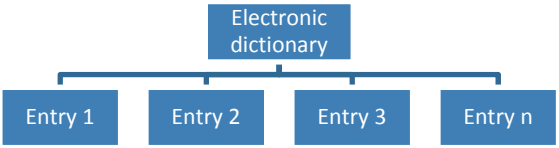

Figure 20: A graphic representation of the architecture of an electronic dictionary

Each entry contains at least a minimum of grammatical information and several of their meanings. (See Figure 21.)

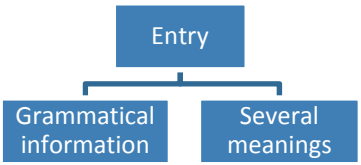

Figure 21: A graphic representation of the structure of an electronic dictionary entry

And then, there is a more the modern approach of using corpora. (See Figure 22.)

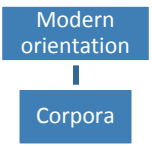

Figure 22: A graphic representation of the modern orientation in lexicological criteria

The architecture of a corpus involves a multitude of entries as well, with there being as many entries as possible. (See Figure 23.)

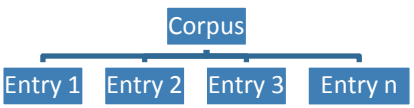

Figure 23: A graphic representation of the architecture of a corpus

However, unlike the electronic dictionaries, each and every entry in a corpus is, in fact, an instance or example of using the same word. (See Figure 24.)

Figure 24: A graphic representation of the content of a corpus entry

Morphological criteria

Morphological criteria focus on the makeup each and every proper word. Essentially, every proper word is made up of a proper root modified by proper affixes. (See Figure 25.)

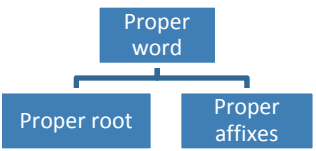

Figure 25: A graphic representation of the make-up of a proper word.

The most basic way of looking of understanding affixes is to classify them according to their position relative to the root. As such there are proper prefixes and proper affixes. (See Figure 26.)

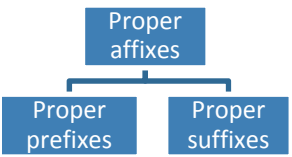

Figure 26: A graphic representation of the representation of the classification of the affixes according to their position

However the user of a translation engine is much more likely to be interested in a rather different kind of distinction as it offers a great deal more information. It is the classification of proper affixes into inflectional affixes and derivational affixes, according to their function. (See Figure 27.)

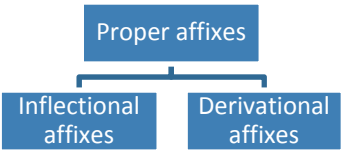

Figure 27: A graphic representation of the classification of the affixes according to their function.

The first and most basic type of properly used affixes, the inflexional affixes, are meant to be added to the root and change the form but not the overall meaning, nor to transform it into another part of speech. It will become very quickly evident for any speaker that inflexional affixes offer information about such features as aspect, case, degree, gender, mood, number, person, polarity, tense. (See Figure 28.)

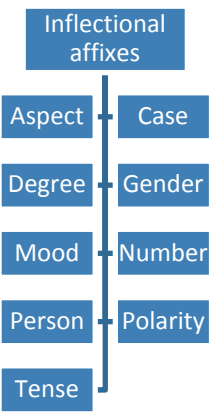

Figure 28: A graphic representation of the features marked by the inflectional affixes

And then there are the derivational affixes. Their number usually far exceeds the number of inflectional affixes, and yet anyone can immediately notice a difference between them. Some derivational affixes only change the overall meaning of the word while others transform a part of speech into an altogether different one. (See Figure 29.)

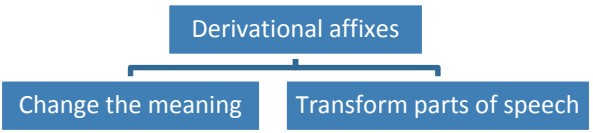

Figure 29: A graphic representation of the types of derivational suffixes

Syntactic criteria

The first of the syntactic criteria is proper word order. Firstly, one is bound to make the distinction between the various statements, namely the affirmative statements, the negative statements and the interrogative statements. (See Figure 30.)

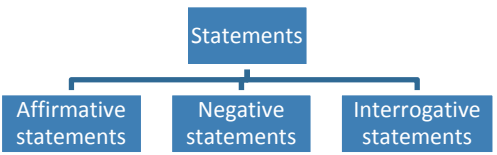

Figure 30: A graphic representation of the types of statements

Secondly, one is advised to know that the basic parts of speech which need be put in the correct order are the subject, the verb and the object. (See Figure 31.)

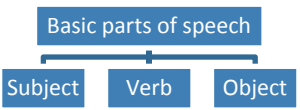

Figure 31: A graphic representation of the basic parts of speech

Thirdly one ought to take a natural language to in order to exemplify the idea of prper word order. For the ease of use it will just have to be English.

Users of translation engines are likely to expect an example of English affirmative statement to be “She likes apples”. It is quite evident that the proper pattern for English affirmative statements is subject-verb-object. (See Figure 32.)

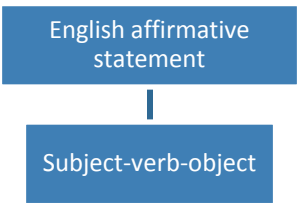

Figure 32: A graphic representation of the proper pattern for English affirmative statements

When it comes to questions, the users of translation engines will probably expect an example of English affirmative statement to be “She doesn’t like apples.” In this case it is safe to say that the proper pattern for English negative statements is subject-verb-verb-object. (See Figure 33.)

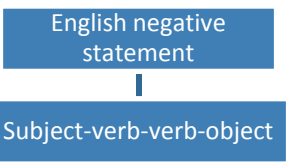

Figure 33: A graphic representation of the proper pattern for English negative statements

However, when confronted with negatives, the users of translation engines are likely to expect an example of English affirmative statement to be “Does she like apples?” Thusly, the proper pattern for English interrogative statements is verb-subject-verb-object. (See Figure 34.)

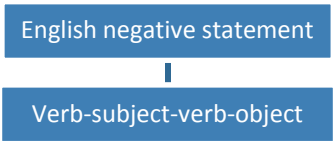

Figure 34: A graphic representation of the proper pattern for English interrogative statements

Armed with this basic knowledge, the user of the translation engine can consider further syntactic criteria. Related to and a logical development of the idea of proper word order is the idea of proper part of speech. According to this idea, a functioning translation engine should have feature enabling it to label or tag parts of speech. (See Figure 35.)

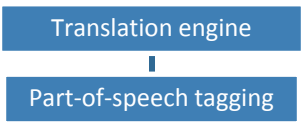

Figure 35: A graphic representation of the Part-of-speech tagging feature of a translation engine.

The underlying thought behind part-of-speech tagging is that it assures the correct use of parts of speech in translation. A well-conceived and well-built Part-of-speech tagger ought to be able to distinguish between and tag nine parts of speech: adjective, adverb, conjunction, determiner, interjection, noun, preposition, pronoun and verb. (See Figure 36.)

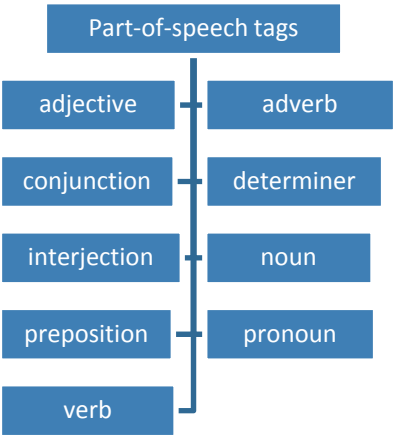

Figure 35: A graphic representation of the Part-of-speech tags.

Conclusion

While this has been nothing but an "essay", quite literally an "attempt" to scratch the surface of the expectancies and perceptions of the users of translation engines, it nonetheless offers a view of the language complexity of the mentioned software applications. This text also shows that there are indeed many limitations to our knowledge of the natural language and even more so to how we mimic it in computational linguistics. In the end, though, it can be seen as a sort of a list of things to do.

Bibliography

Chomsky, N.; Halle, M. [1968]. The Sound Pattern of English. New York: Harper & Row

Harris, J. [1994]. *English Sound Structure*. Oxford: Blackwell

Jespersen, O. [1982]. *Growth and Structure of the English Language*. Chicago and London: University of Chicago Press.

Kroeger, P. [2005]. *Analyzing Grammar: An Introduction*. Cambridge: Cambridge University Press.

Song, J. J. [2012]. *Word order*. Cambridge: Cambridge University Press.

Spencer, A. [1992]. *Morphological Theory*. Oxford: Blackwell.

ÜBER DIE MASCHINELLE ÜBERSETZUNG
(Zusammenfasung)

Schlüsselwörter: *maschinelle Übersetzung, natürlichen Sprache*

Jede maschinelle Übersetzung sollte benutzerfreundlich sein. Somit sollte die Sicht des Benutzers berücksichtigt werden. Dementsprechend sollte jeder Entwickler werden gut informiert über die Struktur menschlichen oder natürlichen Sprache. Dieser Artikel soll alle interessierten Parteien in die Grundlagen der Sprachebenen und ihre Kategorien einzuführen, die für eine erfolgreiche maschinelle Übersetzung relevant sind.

DESPRE TRADUCEREA AUTOMATĂ
(Rezumat)

Cuvinte cheie: *traducere automată, limbaj natural.*

Orice aplicație care permite traducerea automată ar trebui să fie ușor de utilizat. În procesul de dezvoltare a acestor aplicații, ar trebui să se aibă în vedere perspectiva utilizatorului. Dezvoltatorii ar trebui să fie bine informați în ceea ce privește structura limbajului natural. Acest articol are ca scop prezentarea unor elemente introductive în planurile și nivelele limbii precum și ale unor categorii principale ale acestora, care sunt relevante pentru reușita traducerilor automate, tuturor părților interesate.