# Agreement Matters: Challenges of Translating into a Morphologically Rich Language, and the Advantages of a Syntax-Based System

Yoav Goldberg

Department of Computer Science, Ben-Gurion University of the Negev, Israel
yoavg@cs.bgu.ac.il

Consider the following (simple) English sentences: “I drive a car.”, “I don’t know how to drive”, “I wash the car”, “I wash the floor”. Translating them to Hebrew using Google’s statistical MT system, yields:
אני נוהג במכונית (I drive(masculine) a car); אני לא יודעת לנהוג (I don’t know(feminine) how to drive); אני רוחץ את המכונית (I wash(masculine) the car); and אני שוטפת את הרצפה (I wash(feminine) the floor).

While amusing and not quite politically correct, these are all arguably very good translations: without explicit gender marking, the translator can not know if the speaker is masculine or feminine, and he (she?) resorts to deciding based on her (his?) cultural knowledge.

This does, however, highlight a class of problems which arise when attempting to translate from a morphologically clean language (e.g. English) into a morphologically rich one (e.g. Hebrew): many words in the target language are morphologically marked for gender and number, and the translator should be able to generate these markings correctly, based on little, elusive or sometimes no evidence in the source language. These issues are orthogonal to the data sparsity issues associated with highly inflected languages.

Can current state-of-the-art statistical MT systems handle this? In what follows we present a few cases where the target language output should be morphologically marked for either gender or number, with varying amounts (and sources) of information available on the source language text, and discuss the suitability of current translation models to handle these phenomena.

We show that correct handling of morphological agreement is beyond the reach of current systems as it requires better syntactic models, looking beyond a single sentence, and performing accurate anaphora resolution. However, while phrase-based models can not model even the simplest cases, syntax based models already posses most of the necessary machinery.

While we demonstrate using English ⇒ Hebrew translations, similar issues will occur when translating into practically any morphologically rich language. Moreover, the issues discussed remain relevant also when the source language is also morphologically rich.

## 1 Simplest case: explicit agreement.

In Hebrew, pronouns, nouns, adjectives and verbs are morphologically marked for *gender* and *number*. Morphological agreement is required between nouns and their modifiers, between verbs and their subjects, and between coordinated elements. Consider the following sentence pair with their Hebrew translations:

(1a) the committee considered the offer ה וועדה שקלה את ה הצעה

(1b) the organization considered the offer ה ארגון שקל את ה הצעה

In (1a), *committee* is translated to the Hebrew noun וועדה, which is feminine. This requires the verb *considered* to take the feminine form as well (שקלה). In (1b), *organization* correspond to the masculine Hebrew noun ארגון, requiring the verb to take the masculine form שקל. Note that the same verb *considered* is translated as either שקלה or שקל depending on the gender of its subject. Using a feminine verb with a masculine subject is ungrammatical: *ה ארגון שקלה את ה הצעה.

*Can not be handled by a phrase-based system.* This simple case is already beyond the reach of a phrase-based system. Assuming that both [*considered*]-[שקל] and [*considered*]-[שקלה] appear in the phrase-table, the translator should choose the correct one. This role is delegated to the language model, which is likely to prefer וועדה שקלה over וועדה שקל, ensuring grammaticality. Another option is having [*organization considered*]-[וועדה שקלה] in the phrase table as well, and translate them as a single unit. However, both these solutions rely on locality of information, while agreement is a longer distance relation. In our example, the subject NP can be arbitrarily long, e.g. “the committee on solar energy considered . . . ” ⇒ . . . ה וועדה בנושא אנרגיה סולארית שקלה. Here, the subject-head and the verb are separated by three relatively infrequent words, making it practically impossible for an n-gram based language model to come up with an informed decision.

*Syntax-based system* In contrast, translation system that make use of target-side syntax (i.e. string-to-tree systems) are at a much better position to model such longer distant agreement constraints. Consider a system based on xLNTs tree transducer rules (i.e. GHKM rules [4]). Such a system, with rules such as:

| | |
|---|---|
| $\text{NN}_{fem}$(וועדה) $\rightarrow$ committee | *committee* is the feminine noun וועדה |
| $\text{VB}_{fem}$(שקלה) $\rightarrow$ considered | the feminine form of *considered* is שקלה |
| $\text{VB}_{masc}$(שקל) $\rightarrow$ considered | the masculine form of *considered* is שקל |
| S(x0:$\text{NP}_{fem}$ x1:$\text{VP}_{fem}$) $\rightarrow$ x0 x1 | subject-verb agreement (and order) |
| S(x0:$\text{NP}_{masc}$ x1:$\text{VP}_{masc}$) $\rightarrow$ x0 x1 | |

can capture the desired behaviour: once *committee* is translated into a feminine noun, *considered* is forced by the translation rules to take a feminine form as well.

*What is missing.* Of course, a real-world translation system is likely to have over a million translation rules which should be acquired, usually from parsed corpora. Currently, such parsed corpora do not contain the needed morphological annotation for acquiring rules such as those depicted above. Indeed, parsing of morphologically rich languages [7] and in particular modeling agreement [8,5], are still open research questions. While it seems that modeling morphological agreement is only marginally beneficial for parsing accuracy [8,5], such modeling is crucial for accurate translation into a morphologically rich language, making work in line of [6] very appealing.

*Morphologically rich languages on both sides.* It may seem that the problem is easier when both the source and target languages are morphologically rich, as the morphological information is marked also on the source side. This is not the case, because languages differ in their agreement patterns and in the genders they assign to particular nouns. For example, Spanish, like Hebrew, requires Adjective-Noun agreement. However, some nouns (e.g. *computer*) are masculine in Hebrew and feminine in Spanish, and vice-versa. As a result, when translating between Spanish and Hebrew, adjectives are likely to be translated either from feminine to masculine, from masculine to feminine, from masculine to masculine or from feminine to feminine depending on the particular noun they modify. In addition for the need to choose the correct form when translating, which remains as hard as before, acquiring the translation rules becomes harder: the various combinations increase data sparsity, and challenge traditional word-alignment techniques [3,1] which are at the core of most MT systems.

## 2 Harder cases: implicit evidence.

In the previous example, the gender of the noun was known, and the verb followed. This is not always the case. Consider for example "Terry smiled". Here, it is not clear whether the proper name *Terry* is masculine or feminine, and even humans would have trouble translating this correctly without context. For a computer, even a case such as "Margaret smiled" is potentially challenging if the particular proper name was not seen in training and there is no gender information attached to it. This can be alleviated to some extent by acquiring gender preferences for proper nouns automatically from large un-annotated corpora in an unsupervised fashion, as done in [2]. This does not solve the gender-ambiguous names, such as Terry.

*Generating both options.* In case of non-decisive gender ("*Terry smiled and then cried*"), we may want the MT system to propose both masculine and feminine versions. Whether Terry is analyzed as masculine or feminine, both verbs should follow. This is easily accommodated in a syntax based system: a gender decision for one verb will propagate through the syntax and fix the gender of the other verb as well as the gender-unspecified subject. Moreover, gender-preferences knowledge, such as "Terry is 60% likely to be masculine", can easily be incorporated into the model. None of this is possible with a phrase-based system.

*Global inference.* In many cases the gender can be determined based on information in nearby sentences. Consider for example: (2a) *Ms. Elson was arrested. Elson said that . . .* (2b) *Terry smiled. Then he cried.*
In (2a), the first sentence establishes that *Elson* is feminine, forcing *said* in the second sentence to be feminine as well. In (2b), the second sentence indicates that *Terry* is masculine, forcing *smiled* in the first sentence to take the masculine form. Both these cases require anaphora resolution to be performed, either before or jointly with the translation process. Current translation systems do not look beyond a single sentence, and do not attempt to perform anaphora resolution. Translation into a morphologically rich language will require them to do both these things.

## References


[1] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 1993.

[2] Eugene Charniak and Micha Elsner. Em works for pronoun anaphora resolution. In *EACL '09*, pages 148–156, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[3] Hermann Ney Franz Josef Och. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 2003.

[4] Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. What's in a translation rule? In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[5] Yoav Goldberg and Michael Elhadad. Easy-first dependency parsing of Modern Hebrew. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA, 2010.

[6] Reut Tsarfaty. *Relational-Realizational Parsing*. PhD thesis, The Institute for Logic Language and Computation, University of Amsterdam, 2010.

[7] Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. Statistical parsing of morphologically rich languages (spmrl) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Los Angeles, CA, USA, 2010.

[8] Reut Tsarfaty and Khalil Sima'an. Modeling morphosyntactic agreement for constituency-based parsing of modern hebrew. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA, 2010.