MORPHOLOGY, SYNTAX, AND WHAT'S IN BETWEEN: REIFYING CROSS-LINGUISTIC VARIATION THROUGH CROSS-LINGUISTIC PARSING

REUT TSARFATY
UPPSALA UNIVERSITY
REUT.TSARFATY@GMAIL.COM

A central part of NLP research is devoted to the development of parsing systems which analyze the way words combine to form phrases and sentences. This analysis is considered the first step towards natural language understanding, or, as it is often put, extracting "who did what to whom". Such information is crucial for computer programs that perform tasks such as information retrieval, question answering, and machine translation, to name just a few. To date, parsing systems that were developed and applied for English show excellent performance, but the application of the same models to languages such as German, Czech, Arabic, Hebrew, Turkish, Hindi, and more, doesn't necessarily yield comparable results. This discrepancy gives rise to a fascinating challenge, namely: *Is it possible to devise a single parsing system which is abstract enough to accomodate different languages, and yet specific enough to learn from data the structure and properties of particular languages?*

Addressing this question is not only technologically challenging, but is also of utter importance from a scientific point of view. The number of languages in the world is estimated at around 4000-6000 and yet it is striking to realize how similar the different languages are in the principles underlying their organization. In order to formally refer to the commonalities and differences between languages the linguist Noam Chomsky used the term *universal grammar (UG)*, referring to a set of principles which are shared by all languages, and parameters settings that differentiate languages from one another. Modern typological studies following the work of (Sapir 1921, Greenberg 1963) and others play an important part in unraveling the different dimensions of variation between languages (the 'principles') and instantiating their possible values (the parameter 'settings').

Typology tells us, for instance, that all languages have nouns and verbs, and that the relation between a noun and a verb can be of a subject or of an object type. This is the *functional* dimension of language description. Languages vary, however, in how they express these concepts. For instance, subjects in English tend to appear before the verb, but in Warlpiri, subjects can appear almost anywhere in the sentence. This is the *basic word-order* (or, structural) dimension of language variation (Greenberg 1963). Languages also vary in how much information is expressed within a single word. In English, for instance, the subject, verb and object of a sentence are expressed in different words. In Arabic, Hebrew, or Turkish, the subject, verb and object may appear within a single word. This is the *morphological* dimension of variation (Sapir 1921).

It is an indisputable observation that different languages can express the same function differently (Bresnan 2001). For instance, a *subject* relation may be expressed by word order, as in English, or by morphological marking, as in Hebrew and other languages. It has been shown that such variation is what makes the application of state-of-the-art statistical parsing models across languages so challenging (Tsarfaty et al 2010), and that morphological variation is particularly acute for machine translation (this workshop). Nonetheless, this variation has never been empirically quantified based on naturally occurring data. Now, assuming that we can indeed respond to the first challenge by devising a general parsing system that can be applied across language types, and these assign probability distribution to surface structures in the data, another research challenge presents itself, namely: *Is it possible to quantify the variation in the linguistic means that are used to express similar functions in different languages? That is, can we empirically confirm typological conjectures about the surface variation between languages?*

In this talk I present ongoing work that aims to address these two challenges at once. The departure point for the investigation is the Relational-Realizarional parsing architecture of (Tsarfaty 2010) which is based on a set of typological principles, separating the *morphological*, *structural*, and the *functional* dimensions and learning the interplay between them in a recursive fashion. This architecture presupposes word-and-paradigm morphology and extends the paradigmatic view to the domain of syntax, thus obtaining a coherent way to display the interface between syntax and morphology in realizing grammatical relations in different languages. While the overall paradigmatic structure of the architecture reflects universal language organization, we have been able to successfully apply it to two typologically different languages, **Hebrew**, a Semitic languages, and **Swedish**, a Germanic language. In both languages are able to show, so far for gold standard input only, that the incorporation of morphological information leads to improved performance, albeit exploiting different argument marking patterns. Furthermore, the empirical distributions learned for the same parameter classes in the two different languages reflect the variation in the linguistic structure of the two input languages.

Based on these results we hypothesize that, at least in principle, it is possible to devise a general model based on universal organizational principles and learn cross-linguistic variation from data, without changing the design of the architecture. Assuming that this application of the architecture can yield comparable parsing results across languages, we may then view the logical structure of the learner as defining the 'principles' that govern all human languages. The set of model statistical distributions of the parameters in different corpora then empirically reflect different parameter settings as distribution over typological parameter values ('parameter settings'). We conjecture that applying this architecture to parsing a greater set of languages will allow us to refine our understanding of cross-linguistic variation by assigning quantitative measures to it. This, in turn, has the potential of aiding the development of better and more accurate machine translation systems, by making use of the knowledge of how the source and target languages vary in their structure, and exploiting the statistical reflection of this variation in the data.