# *A resource-light approach to morpho-syntactic tagging*
*Anna Feldman and Jirka Hana*

This volume deals with the problem of morpho-syntactic tagging, i.e. "the process of assigning part-of-speech, case, number, gender and other morphological information to each word in a corpus" (p. 2). In its simplest form, a part-of-speech tagger identifies the traditional grammatical categories such as verb, noun, adverb, adjective, preposition, and so on.
In the case of morphologically rich languages, the problems of part-of-speech tagging and morphological analysis are closely tied, and both tasks are often integrated in one program.

In the domain of natural language processing, part-of-speech tagging is regarded as a first level of abstraction in text analysis and is often used as a pre-processing module in many language technology applications such as parsing, information retrieval, spelling error correction, speech synthesis, and text mining (Daelemans and van den Bosch, 2005: 86-87). Morpho-syntactic taggers thus form an indispensible resource for the development of a wide range of NLP applications.

Most standard tagging algorithms are corpus-based, which means that they acquire the information they use during the tagging process (e.g. tag frequencies, tag sequence probabilities, rule sets) from pre-annotated corpora. They thus heavily rely on the existence of high-quality annotated training data. As the creation of such resources is time-consuming, such corpora do not exist for all languages.

This volume describes an alternative approach to morpho-syntactic tagging by porting the relevant information from one language to a related language, avoiding thus the labour-intensive creation of an annotated corpus. The approach has been tested on two morphologically rich language families: a Slavic language pair (porting from Czech to Russian) and two Romance language pairs (porting from Spanish to Catalan and from Spanish to Portuguese).

The volume is composed of eight chapters: Introduction; Common tagging techniques; Previous resource-light approaches to NLP; Languages, corpora and tag sets; Quantifying language properties; Resource-light morphological analysis; Cross-language morphological tagging; Summary and further work. The volume is to a large extent based on the doctoral dissertation of Anna Feldman (2006).

Chapter 1 explains in more detail what the book offers and explains how the book is organized.

Chapter 2 describes commonly used techniques for building part-of-speech taggers. In the description, an important distinction is made between supervised and unsupervised methods. Supervised methods rely on pre-annotated corpora as training data, while unsupervised methods use untagged corpora and automatically induce the tag set. The appropriateness of different tagging techniques for highly inflected languages is discussed at the end of this chapter.

Chapter 3 describes two types of resource-light approaches to NLP tasks: unsupervised or minimally supervised methods and methods that use cross-lingual knowledge transfer. The latter methods use the resources that are available in one language to project linguistic knowledge into a related resource-poor language.

Chapter 4 focuses on the resources that were used in the experiments described in chapter 6 and 7. The chapter starts off with a discussion of the morpho-syntactic properties of Czech, Russian, Spanish, Catalan and Portuguese, and continues with a description of the different types of corpora and tag sets. Special attention is given to the benefits of positional tag systems.

In chapter 5, some properties of the Slavic and Romance languages under examination and their tag sets are investigated. The chapter also touches upon the issue of tag set reduction.

Chapters 6 and 7 present the core part of the volume. Chapter 6 introduces a resource-light approach to morphological analysis and chapter 7 describes a number of experiments in cross-language morphological annotation transfer from Czech to Russian and from Spanish to Catalan and Portuguese.

Finally, chapter 8 summarizes the work and describes some further research paths that

can be explored.

The volume presents a new approach for the rapid development of morpho-syntactic taggers for languages for which no annotated corpora are available. The method presented in this volume relies on un-annotated corpora and available resources for related languages.

One of the merits of this volume is that it focuses on languages other than English, the dominant language in the research area of natural language processing. More specifically, it focuses on the problem of tagging morphologically rich languages.

Another asset of this volume is that it touches upon some interesting issues such as language typology, the impact of tag set size on tagging accuracy, and tag set standardization. Especially the last point is an important issue in the development of multilingual applications (Steinberger 2010).

Nevertheless, a few critical observations can be made concerning the volume. The description of the common tagging techniques (chapter 2) stands fairly separate from the experiments presented in chapters 6 and 7 and the text fails to explain why certain decisions have been. Moreover, it is not entirely clear to what extent the cross-lingual porting is a manual or an automatic process, which e.g. utilizes machine learning techniques.

Another critical remark is that the authors fail to explain important concepts and metrics used throughout chapters 6 and 7. For example, the concepts *emissions* and *transitions* are not adequately explained in the text; different evaluation metrics are used (*error rate, recall error, accuracy*), but the authors do not describe how these metrics are related to each other.

A minor disadvantage is that the volume contains some editing errors, which careful proofreading should have caught (e.g. the table with entropy calculations and the table with mutual information calculations are missing in the text; on p. 69 the sentence referring to the description of corpora in the next section is obsolete).

In summary, I found the topic of the volume very interesting as it broadens the scope of the research field of morpo-syntactic tagging to under-resourced morphologically rich languages. For researchers planning to rapidly build new morpho-syntactic taggers for Slavic and Romance languages, the book is certainly a valuable resource.

As the authors mention, the book touches upon a number of topics such as typology, morphology, linguistic annotation, cross-lingual studies and applications and aims at researchers and students who are interested in these scientific areas. However, the volume might be too superficial and in some cases too detailed to be of value to a broad audience.

References

**Feldman, A.** (2006). *Portable Language Technology: A Resource-light Approach to Morphosyntactic tagging*. Ph.D. Dissertation. The Ohio State University.

**Daelemans, W. and van den Bosch, A.** (2005). *Memory-based language processing*. Cambridge University Press, Cambridge, United Kingdom

**Steinberger, R.** (2010). *Challenges and methods for multilingual text mining*. Proceedings of the Seventh Language Resources and Evaluation Conference (LREC 2010). Keynote speech.