# Issues of Theoretical Linguistics in Natural Language Processing and Computational Linguistics

## Prof. Matteo Pellegrini

***COURSE AIMS AND INTENDED LEARNING OUTCOMES***

The course aims at introducing students to advanced notions of theoretical linguistics, showing on the one hand how such notions can help to create more accurate lexical and textual resources and more fine-tuned tools, on the other hand how NLP techniques allow to capture aspects of language structure that are sometimes neglected in purely theoretical approaches.

At the end of the course, students will have a solid knowledge of fundamental notions employed in the description of the morphology of natural languages, being aware of the different theoretical perspectives that can be taken. They will thus be able to critically examine the tagsets used for the annotation of corpora and other resources and the linguistic features exploited by NLP tools, understanding the theoretical assumptions presupposed by different schemes and approaches. Furthermore, they will have become familiar with tools and resources that allow to perform useful morphological analyses on language data in an automatic way.

***COURSE CONTENT***

The course is divided into two parts.

The first part consists of lectures given by the teacher on selected advanced topics of theoretical morphology, with special attention to quantitative methods and their application in electronic resources and tools. More specifically, the topics addressed include the following:

* distinction between lexemes and wordforms;
* morphological processes: affixation, compounding, conversion, etc.;
* inflection, derivation and the syntax-morphology interface;
* inflectional paradigms: theoretical issues and computational approaches, the reinflection task;
* productivity in word-formation: definition and corpus-based measures;
* theoretical debate on morphology: morpheme-based and word-based frameworks, Item-and-Arrangement, Item-and-Process, Word-and-Paradigm models, constructive and abstractive approaches;
* resources and tools for computational morphology: inflected lexicons (Unimorph project); the Principal Part Analyzer; the Qumin toolkit for the quantitative modelling of morphology.

In the second part of the course, students will be invited to present their own work, building on the notions they acquired in the first part. They can choose whether to focus on a more specific topic in theoretical morphology, based on additional readings, or to discuss the results of the application of automatic tools to some new data. In both cases, the content of the presentation will have to be agreed upon with the teacher in advance.

***READING LIST***

The main reference handbook is:

Haspelmath, M. & Sims, A. D. 2011. *Understanding Morphology*. 2nd edition. London: Hodder Education.

Further information can be found in the following advanced handbooks, from which selected chapters will be taken as teaching support:

Audring, J. & Masini, F. (eds.). 2019. *The Oxford Handbook of Morphological Theory*. Oxford: Oxford University Press.

Baerman, M. (ed.). 2015. *The Oxford Handbook of Inflection*. Oxford: Oxford University Press.

Hippisley, A. & Stump, G. (eds.). 2016. *The Cambridge Handbook of Morphology*. Cambridge: Cambridge University Press.

Lieber, R. & Štekauer, P. (eds.). 2011. *The Oxford Handbook of Compounding*. Oxford: Oxford University Press.

Lieber, R, & Štekauer, P. (eds.). 2014. *The Oxford Handbook of Derivational Morphology*. Oxford: Oxford University Press.

Additional readings will be indicated during the course, also on the basis of the students’ interests and inclinations. A non-exhaustive and non-mandatory list is provided below:

Ackerman, F. & Blevins, J. P. & Malof, R. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In Blevins, J. P. & Blevins, J. (eds.). *Analogy in grammar: Form and acquisition*. Oxford: Oxford University Press, 54–82.

Baayen, R. H. 1992. Quantitative aspects of morphological productivity. In Booij, G. & van Marle, J. (eds.). *Yearbook of Morphology 1991*. Dordrecht: Kluwer, 109–149.

Baayen, R. H. 1993. On frequency, transparency and productivity. In Booij, G. & van Marle, J (eds.). *Yearbook of morphology 1992*. Dordrecht: Kluwer, 181–208.

Baayen, R. H. & Lieber, R. 1991. Productivity and English derivation: a corpus-based study. *Linguistics* 29, 801–43.

Beniamine, S. & Bonami, O. & Sagot, B. 2017. Inferring inflection classes with description length. *Journal of Language Modelling* 5(3), 465–525.

Blevins, J. P. (2006). Word-based morphology. *Journal of linguistics*, 42(3), 531–573.

Bonami, O. & Beniamine, S. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2), 156–182.

Bonami, O. & Boyé, G. 2014. De formes en thèmes. In Villoing, F. & Leroy, S. & David, S. (eds.), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*. Paris: Presses Universitaires de Paris-Ouest, 17–45.

Booij, G. 1993. Against split morphology. In Booij, G. & van Marle, J. (eds.). *Yearbook of Morphology 1993*. Dordrecht: Kluwer, 27–49.

Cotterell, R. & Kirov, C. & Sylak-Glassman, J. & Yarowsky, D. & Eisner, J. & Hulden, M. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 10–22.

Finkel, R. 2016. Computer-Based Tools for Word and Paradigm Computational Morphology. In *Oxford Research Encyclopedia of Linguistics.*

Finkel, R. & Stump, G. 2007. Principal parts and morphological typology. *Morphology* 17, 39–75.

Gaeta, L. & Ricca, D. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1), 57–89.

***TEACHING METHOD***

Lectures (in English) with occasional exercises and use and demonstration of software in class in the first part of the course; presentations by the students and related questions and discussion in the second part of the course.

***ASSESSMENT METHOD AND CRITERIA***

During the final part of the course, there will be an evaluation of the work presented by the student, based on:

the understanding of the topics of the presentation;

the quality of the supporting material (handout or slides);

the clarity and effectiveness of the exposition.

After the course, there will be an oral exam to verify the acquisition of the contents of the course. The evaluation will be based on:

the knowledge of the topics taught in the course;

the ability to apply the notions to new data;

the appropriate use of the technical terminology of the discipline.

There will be a single final grade, based both on the evaluation of the presentation (50%) and on the outcome of the oral exam (50%).

***NOTES AND PREREQUISITES***

Given the advanced nature of the course, knowledge of basic notions of theoretical linguistics – the ones that are taught in any introductory course – is required. Students that fail to meet this prerequisite should contact the teacher to discuss their situation and establish a list of readings to bridge this gap before the beginning of the course.

*Place and time of consultation hours*

Place: CIRCSE Research Center, Franciscanum building, second floor, room n. 215.

Time: on appointment, by sending an e-mail to matteo.pellegrini@unicatt.it.