UniDive Training Summer School 2024

Chiara Di Maio

Title: An Annotated Corpus for Buranello

Name and Affiliation: Chiara Di Maio, CUNY Graduate Center (U.S.)
(Italian citizen)

Keywords: Universal Dependencies, low-resource language, Italian, dialect.

Description of a resource related to the topics of the training school:
My project would involve the use of the Universal Dependencies framework for consistent annotation of the grammar of a low-resource language.

Explanation of how the participation in the training school will be useful for the project:
My interest in attending the *UniDive Training Summer School* is tightly connected to a long-term research project that I am currently planning as a first-year Ph.D. student. The project consists in creating an annotated corpus for Buranello, of which I am a heritage speaker.

Buranello is a language spoken in Burano, an island which counts approximately 2,200 residents and is located in north-eastern Italy. This language is under-studied and under-documented and is considered vulnerable (only spoken in a limited number of domains) by the UNESCO Atlas of the World's Languages in Danger (under the name 'Venetan', which includes different varieties spoken in the same region). The grammar of Buranello has not yet been thoroughly described and analyzed. In fact, an extremely limited amount of work has been done on the language.

One of my main goals is to create different resources for Buranello, including a corpus (which, to the best of my knowledge, has not yet been created). Since there is no annotated corpus of any kind for this language, there also is none that employs a framework such as Universal Dependencies to annotate grammar consistently. I believe I could develop such project to its full potential thanks to my academic background and my personal experience. For instance, I have already been working on various theoretical aspects of the syntax and morphology of Buranello and Italian for a few years now. Such studies certainly constitute an advantage in terms of annotation, as I have already analyzed some of the syntactic structures and morphological features of the language. Furthermore, as a heritage speaker of Buranello, I have some intuition on its grammar, which is supported by the knowledge I gained from my formal training in linguistics. Given the island's low number of residents and the endangerment status of the language, there might not be many researchers who are both speakers of the language and trained linguists. Therefore, my participation in the *UniDive Training Summer School* would probably be a unique opportunity to carry out this work.

For these reasons, I believe gaining the skills and practical experience at the *UniDive Training Summer School* early on in my graduate career would be a great opportunity for me to maximize the potential for my project. In particular, all activities proposed – *Annotation of*

Universal Dependencies treebank for a new language, *Annotation of multiword expressions in a new language*, and *Corpus annotation infrastructure* – seem to fit very well with the goals I set for my project, and I am sure they will be extremely useful through each step of its development.

Open questions related to the project which could be addressed during the brainstorming hackathon:

- Buranello is a language that has no official orthography. What is the most practical way to transcribe it in preparation for the creation of an annotated corpus? Should I develop a practical orthography, for instance modeling it off the basic conventions of Italian or other nearby Romance languages, or should I use the IPA?
- Since it is un under-documented language, the grammar of Buranello has not been thoroughly described yet. How can annotation of grammar, including POS tagging, morphological features, and syntactic dependencies be approached when the grammar of a language has not yet been analyzed?
- At what point of the creation of a corpus should multiword expressions annotation be implemented? What would be more advisable? Starting right away or annotating them later?

Short statement of the project phase (planning, started, in the process of creation):
As a first-year Ph.D. student, I am currently planning this research project. I will start collecting data this summer and continue to develop it throughout my studies. By the end of my second year, I will have transcribed my data and started the creation of the corpus. My third and fourth year will be devoted to annotation, and in my fifth year I will be able to query my corpus to answer some research questions. For instance, Buranello exhibits two different structures for interrogative clauses, subject-verb or verb-subject. Does one structure occur more frequently than the other? Are there any other environments where this inversion occurs? As previously mentioned, implementing multiword expressions annotation would also be a possible further expansion of the corpus. These are only some of the questions that could be answered once an annotated corpus has been created. Another advantage of making such resources available for under-studied languages is that they can also help explain existing theories or even develop new ones.

References

Moseley, Christopher (ed.). 2010. Atlas of the World's Languages in Danger, 3rd edn. Paris, UNESCO Publishing. Online version: http://www.unesco.org/culture/en/endangeredlanguages/atlas

Nivre, Joakim, Zeman, Daniel, Ginter, Filip, and Tyers, Francis. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.