CLARIN

Common Language Resources and
Technology Infrastructure

Tour de CLARIN
HUNGARY

Edited by **Darja Fišer** and **Jakob Lenardič**

HUNGARY

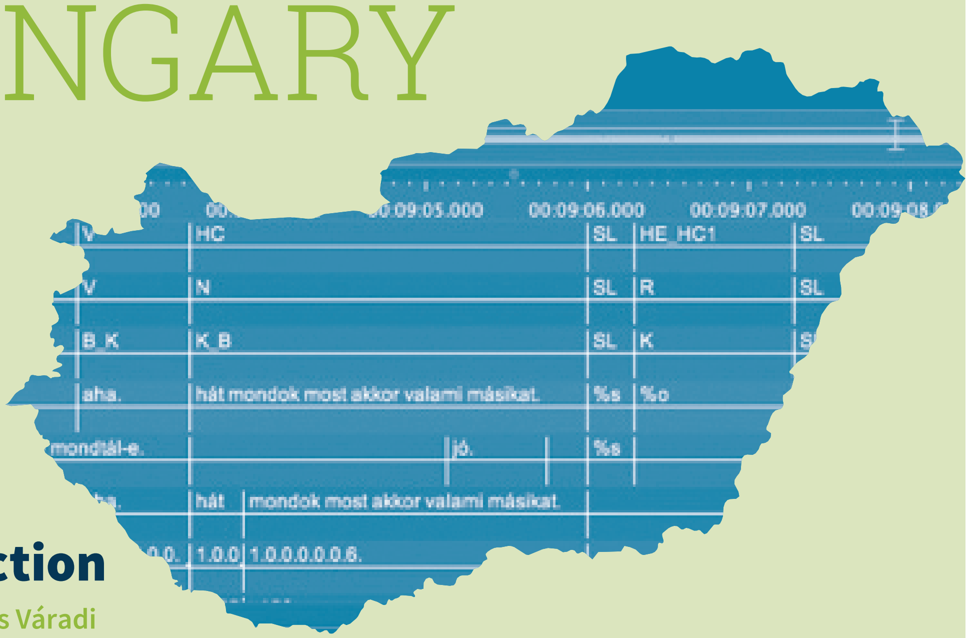
Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia and knowledge centres as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

The brochure presents Hungary and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium’s infrastructure in their research

HUNGARY



Introduction

Written by **Tamás Váradi**

The national CLARIN consortium for Hungary, HUN-CLARIN, joined CLARIN-ERIC in 2016.[1] The Research Institute for Linguistics was one of the founding partners of CLARIN and took an active role in the preparatory phase of the history of CLARIN. The members of the consortium are the Research Institute for Linguistics, the MOKK Centre for Media Research and Education and the Speech Communication and Smart Interactions Laboratories of the Budapest University of Technology and Economics, the University of Szeged, the University of Debrecen, the Pázmány Péter Catholic University, the MorphoLogic LLC, the Institute for Computer Science and Control, and the Institute of Cognitive Neuroscience and Psychology. The national coordinator for HUN-CLARIN is Tamás Váradi.

As can be seen from the above list, the consortium covers a wide range of complementary expertise and research interests. It represents most of the leading research centres in Hungarian language and speech technology, which have closely cooperated in various national and international projects for more than a decade.

The resources developed by HUN-CLARIN members include corpora that are indispensable to research in the use of the Hungarian language, such as the Hungarian National Corpus, which has recently been upscaled to giga size, the Hungarian WebCorpus, which was the first of its kind in Hungarian, and the Szeged Treebank, the reference treebank for Hungarian. Bilingual resources include the Hunglish Corpus, a sentence-aligned Hungarian-English parallel corpus of about 120 million words in four million sentence pairs. A truly unique resource is HuComTech Corpus, a large scale multimodal corpus which offers a rich dataset on 47 annotation levels and was presented to the CLARIN community at the CLARIN 2018 Conference.

As regards tools, the Hun* set of tools developed by the MOKK Centre (such as HunAlign, HunTag, HunMorph, etc.) has also acquired recognition beyond Hungary for its versatility and free availability for languages other than Hungarian. A major recent achievement is the comprehensive processing chain e-magyar, which was developed through widespread collaboration within HUN-CLARIN members. This open and modular toolset was developed to suit the needs of Digital Humanities researchers and application developers alike, and is therefore available both as a web service and for download from GitHub repositories.

Severely limited by lack of funding for national activities, HUN-CLARIN, nevertheless, is making successful efforts to reach out to the Humanities and Social Science communities. It has established cooperation with the Centre for Digital Humanities at Eötvös Loránd University as well as the Centre for Social Sciences. Last year HUN-CLARIN embarked on a roadshow among Hungarian universities showcasing the central HUN-CLARIN tools and resources as well as local research projects. The three events so far at the universities of Szeged, Debrecen and Pécs have proved so popular that a second event is already being organized this autumn at Szeged University, at the institution's request.

In 2017 HUN-CLARIN hosted the CLARIN Annual Conference in Budapest. In the future, HUN-CLARIN plans to establish a K-Centre for Hungarian, continue with our outreach efforts and, subject to securing some national funding, set-up and operate a B-Centre as well.

The HUN-CLARIN team

[1] https://clarin.hu/en

Tool | **e-magyar: a Comprehensive Processing Chain for Hungarian**

Written by **Balázs Indig** and **Tamás Váradi**

The e-magyar toolchain was developed in 2016 as a major collaborative effort across the Hungarian NLP community.[2] The rationale for it was based on a clear vision of an open, modular, extendable and easy-to-use pipeline for Hungarian, which was suitable for non-specialists and developers alike. There existed pipelines created especially for Hungarian (e.g. the Hun* tools or Magyarlánc), and state-of-the art pipelines (e.g. StanfordNLP and UDPipe) also support Hungarian. However, they cannot fulfil the desired functions of modularity, extendibility and user-friendliness. For example, improving the existing methods and annotations on different levels of processing was extremely tedious, which prematurely cut short almost every attempt at natural improvement.

Therefore, the development of e-magyar started by collecting and integrating the best-practices and good features of the existing modules and pipelines while implementing the features that the community missed the most. The first version was integrated into the GATE framework.

The toolchain consists of the following tools (see Figure 1 for the general architecture):

- emToken, a rule-based tokenizer which adds Unicode handling and detokenization to its ancestor Huntoken;
- emMorph, a rule-based morphological analyser based on Helsinki Finite State Transducer, the flagship tool within e-magyar which integrates all previous efforts (including the commercial tool HUMOR) into a new, open-source tool for Hungarian;
- emPOS, a statistical PoS-tagger derived from HunPOS which is an improved version of the TnT tagger;
- emDEP, a dependency parser and emCONS a constituent parser taken directly from Magyarlánc;
- emNER, a named entity recognizer based on the HUNtag3 framework;
- emChunk, a NP recognizer based on the HUNtag3 framework.

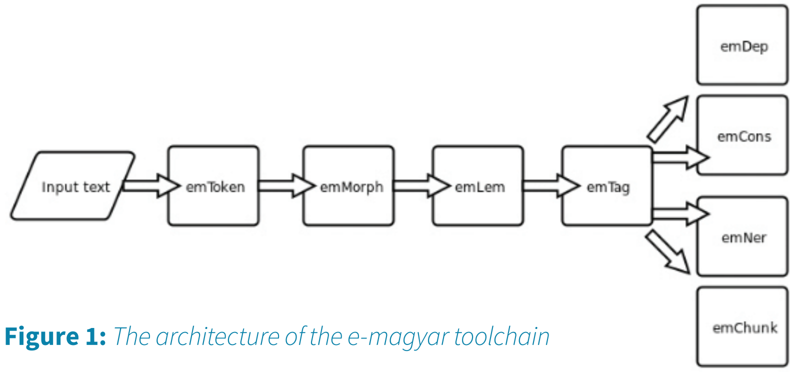
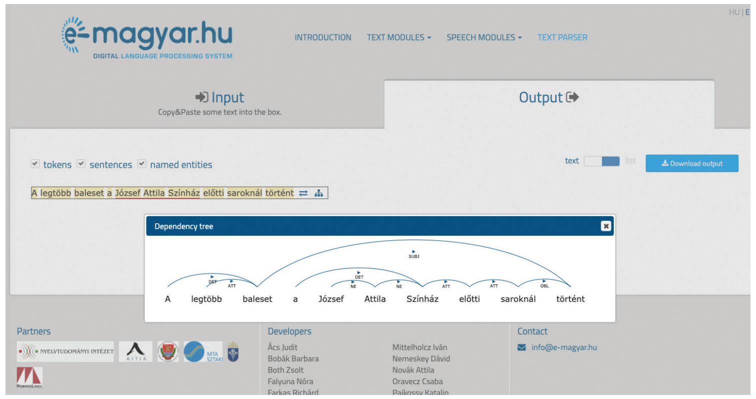


Figure 1: *The architecture of the e-magyar toolchain*

To further improve the efficiency and user-friendliness of e-magyar, the whole architecture was given a thorough overhaul in which the GATE framework was replaced with an inter-module communication framework that follows the toolbox philosophy. The new architecture makes e-magyar not only a truly modular, easy-to-use and extendable toolchain, but one that can very quickly be transformed into a webservice and a Python library as well. To illustrate the modularity and enhanced flexibility of the system, many new modules have already become part of the toolchain, providing alternative options to existing modules. For example, the well-known spellchecker and stemmer Hunspell presents an alternative to emMorph and the three UDPipe modules – tokenizer, PoS-tagger, dependency parser can be selected in preference to emToken, emPOS and emDEP.

The e-magyar toolchain was developed to suit non-technical users as well. They can use the drag-and-drop Text Parser webservice which accepts short texts and outputs their analysed version to the selected level of detail (see Figure 2). In addition, a more lightweight web option, a web service of emMorph, (showing the morphological analysis of individual words) was also set up to enable linguists to check the analyses of particular words during their annotation work.



A legtöbb baleset a József Attila Színház előtti saroknál történt

Ács Judit
Bobák Barbara
Both Zsolt
Falyuna Nóra
Farkas Richárd
Mittelholcz Iván
Nemeskey Dávid
Novák Attila
Oravecz Csaba
Pajkossy Katalin

info@e-magyar.hu

Figure 2: *Dependency parsing in the Text Parser webservice*

[2] http://e-magyar.hu/en

Resource | **Multimodal HuComTech Corpus**

Written by **László Hunyadi** and **Tamás Váradi**

The idea of building a multimodal corpus of Hungarian (containing annotations of text, prosody, gaze, gesture, etc.) was conceived 10 years ago. The aim was to improve human-machine communication applications (like chatbots) by empowering them with a comprehensive set of knowledge about human-human communicative behaviour. The underlying assumption was that there exist certain primitives of human behaviour. Such behavioural primitives form temporal patterns which can be assigned functional interpretations. For instance, a prosodic feature like falling intonation followed by a visual cue such as a downward gaze often signals that the speaker wishes to terminate his or her turn in conversation. Such primitives can further serve as a marker which, with a certain probability, points to a pattern with a given interpretation.

When building the HuComTech corpus[3] we first observed and annotated primitives of behaviour at multiple levels, which included recording intonation, morpho-syntactic annotation, video annotation, unimodal and multimodal pragmatic annotation, among others. Subsequently, we interpreted the complex raw annotation phenomena in terms of pragmatic and communicative function, and finally we identified actual patterns of behaviour based on the annotated raw and interpreted data. In total, about 50 hours of dialogues with 111 subjects were recorded in two (formal and informal) scenarios. HunCLARIN experts captured the multimodality of human-human communication by observing a wide range of both non-verbal and verbal behaviour. The primitives of non-verbal behaviour were either visual or audio in nature. The visual primitives included eye gaze (direction and blinking), eyebrows, head, hand and (upper) body movement, perceived emotions and a range of pragmatic and communicative categories (such as turn management, agreement, certainty, etc.). The non-verbal audio primitives included a range of prosodic features, perceived emotions and a range of pragmatic and communicative categories.

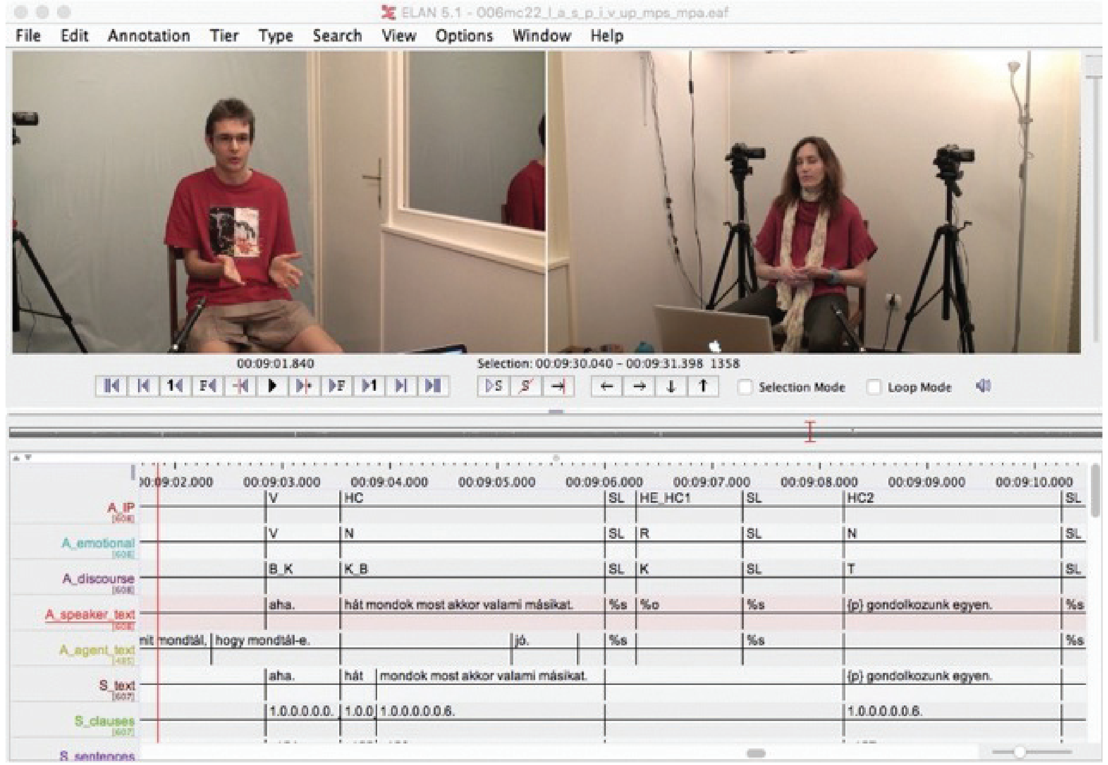
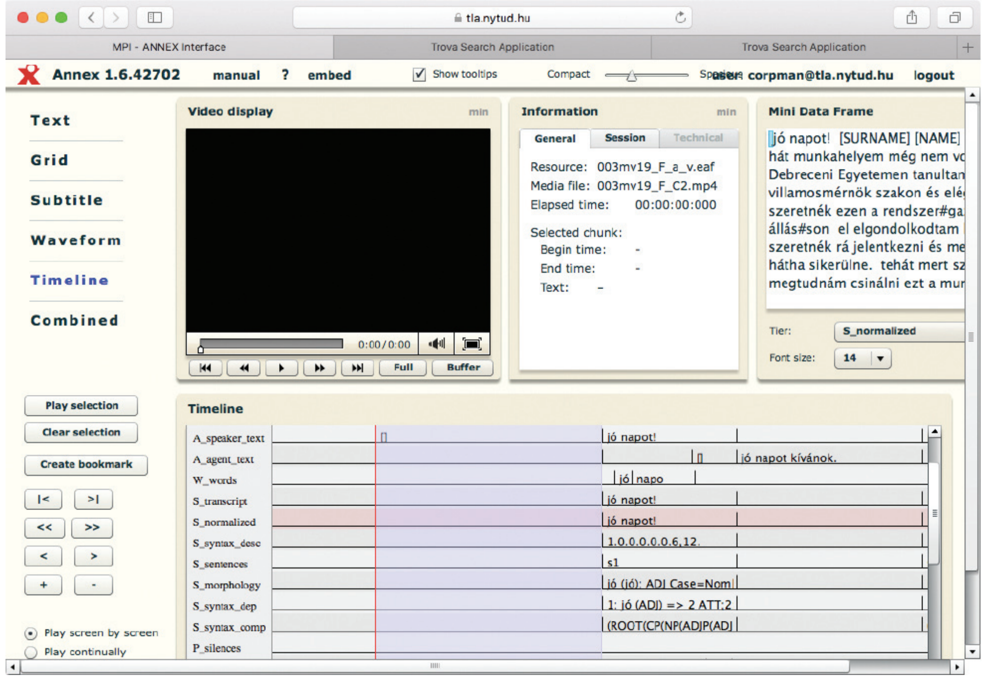

ELAN 5.1 - 006mc22_l_a_s_p_i_v_up_mps_mpa.eaf
File Edit Annotation Tier Type Search View Options Window Help
00:09:01.840
Selection: 00:09:30.040 - 00:09:31.398 1358
□ Selection Mode
□ Loop Mode
00:09:02.000 00:09:03.000 00:09:04.000 00:09:05.000 00:09:06.000 00:09:07.000 00:09:08.000 00:09:09.000 00:09:10.000
A_IP: V | HC | SL | HE_HC1 | SL | HC2 | SL
A_emotional: V | N | SL | R | SL | N | SL
A_discourse: B_K | K_B | SL | K | SL | T | SL
A_speaker_text: aha. | hát mondok most akkor valami másikat. | %s | %o | %s | {p} gondolkozunk egyen. | %s
A_agent_text: nit mondtál, | hogy mondtál-e. | jó. | %s | %s | %s
S_text: aha. | hát | mondok most akkor valami másikat. | {p} gondolkozunk egyen.
S_clauses: 1.0.0.0.0.0. | 1.0.0 | 1.0.0.0.0.0.6. | 1.0.0.0.0.0.6.
S_sentences

Figure 3: *Annotation of the HuComTech corpus with the ELAN tool*

The annotation of verbal primitives was aimed at offering the first of its kind of syntactic analysis of spoken language. This was partly done by using another Hungarian CLARIN tool, *magyarlanc*, while the German CLARIN *webmaus* tool was used for the alignment of words on the timeline. The specific features of the HuComTech Corpus include its unique conception of multimodality, which actually represents the synthesis of three approaches: the annotation of primitives and functions both based on visual observation alone, the same kind of annotation based only on audio observation, and genuine multimodality based on audio and visual clues together. This threefold distinction of primitives within multimodality allows for capturing behavioural patterns at three levels (vision, audio, and their joint complexity), facilitating the building of two-way communication systems. The corpus has also been successfully used in linguistic research; for instance, Hunyadi (2019) used HuComTech to study the multimodal expression of agreement and disagreement, while Szekrényes (2019) presented an approach to the post-processing of temporal patterns based on the multimodal data in the corpus.

[3] http://tla.nytud.hu/



Resource: 003mv19_F_a_v.eaf
Media file: 003mv19_F_C2.mp4
Elapsed time: 00:00:00:000

Selected chunk:
Begin time: -
End time: -
Text: -

jó napot! [SURNAME] [NAME]
hát munkahelyem még nem vo
Debreceni Egyetemen tanultan
villamosmérnök szakon és elé
szeretnék ezen a rendszer#ga
állás#son el elgondolkodtam
szeretnék rá jelentkezni és me
hátha sikerülne. tehát mert sz
megtudnám csinálni ezt a mur

Tier: S_normalized
Font size: 14

A_speaker_text	jó napot!
A_agent_text	jó napot kívánok.
W_words	jó napo
S_transcript	jó napot!
S_normalized	jó napot!
S_syntax_desc	1.0.0.0.0.0.6,12.
S_sentences	s1
S_morphology	jó (jó): ADJ Case=Nom
S_syntax_dep	1: jó (ADJ) => 2 ATT:2
S_syntax_comp	(ROOT(CP(NP(ADJP(ADJ
P_silences	

Figure 4: *Browsing the corpus at the HUN-CLARIN repository tla.nytud.hu with the Annex tool*

For more information about the HuComTech corpus see Hunyadi et al. (2018).

References:

Hunyadi, L. 2019. Agreeing/Disagreeing in a Dialogue: Multimodal Patterns of Its Expression. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.01373.

Hunyadi, L. et al. 2018. Human-human, human-machine communication: on the HuComTech multimodal corpus. In *Proceedings of the CLARIN Annual Conference* 2018, 56–65.

Szekrényes, I. 2019. Post-processing T-patterns Using External Tools From a Mixed Method Perspective. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.01680.

Event | **The HUN-CLARIN Roadshows**

Written by **Réka Dodé**

HUN-CLARIN, in line with the overall CLARIN mission, considers it a high priority to support humanities and social science research with suitable language technologies and language resources.

Unfortunately, Digital Humanities in Hungary has recently suffered a serious setback when the budding DH training at the master's level was suddenly stopped by the Hungarian Government. While this did not mean an end to Digital Humanities research at the Centre for Digital Humanities – Eötvös Loránd University, with which HUN-CLARIN have established good cooperation, it did certainly make the task of reaching out to DH researchers a lot more challenging.
To counter this difficulty, HUN-CLARIN devised the concept of a Roadshow series that is based on the idea of proactively (and literally) bringing language technology to where humanities research is actually done, namely, to Hungarian universities. The other key concept behind the Roadshows is that instead of a one-sided evangelization of language technology, the workshop should mobilize and showcase local initiatives.

The template for the workshops is therefore divided into two parts. The first part features an introduction to HUN-CLARIN and the value proposition of CLARIN in general. The introduction is then followed by an overview of the basic resources and tools HUN-CLARIN offers to Humanities researchers. The second part highlights local Humanities projects where there had already been an initial attempt to employ some language technology tool in research, or where an interesting dataset and/or a vaguely perceived need for some automated method for analysing the data had been identified. Both parts of the workshops have been designed for plenty of interaction between the visiting HUN-CLARIN team and the hosting researchers, but the afternoon sessions in particular are buzzing with excitement for both teams.

The visiting HUN-CLARIN team is fascinated by the genuinely interesting research questions and the data that the local colleagues have shown in their presentations. Researchers of the host institution, on the other hand, appreciate the advice they have received and the perspectives that it has opened with regard to their own work. Because the research profiles of the hosting institutes have been different, each event has had a different focus and offered intellectual excitement of a different kind, but with a consistently high level of intensity.
So far HUN-CLARIN has organized three such Roadshow events. The first took place at Szeged

University on 18 October 2018 and was held with the support of a grant by CLARIN ERIC.[4] The second workshop was staged at Debrecen University on 7 February 2019,[5] and the third event was organized in Pécs on 2 May 2019.

In Szeged the local contributions focussed on speech and language technology tools, in Debrecen the topics of the local presentations centred on questions of building and using the HuComTech multimedia corpus, while in Pécs the discussion revolved around computational linguistic approaches to discourse representation.

HUN-CLARIN plans to take the Roadshow to other universities in the future. We take encouragement by the fact the local organizers of the Szeged workshop have since indicated that in response to local demand they want to stage a follow-up workshop in November 2019.

The HUN-CLARIN Roadshows

[4] http://clarin.hu/en/content/seminar-speech-and-language-technology-tools-0

[5] https://clarin.hu/en/content/use-corpora-language-technology-tools-and-data-driven-methods-human-sciences

Interview | **Noémi Vadász**

Noémi Vadász is a PhD student and junior research fellow who works at the Research Institute for Linguistics. As a computational linguistic with a formal background in syntax and semantics, she collaborates with HUN-CLARIN in the e-magyar project.

Please describe your academic background and your current research position.

<

I am a junior research fellow at the Research Institute for Linguistics, Hungarian Academy of Sciences at the Research Group of Language Technologies. After my BA on Hungarian Literature and Linguistics I have finished two MA programmes: Theoretical Linguistics and Computational Linguistics. I then moved on to the Doctoral School for Linguistics at Pázmány Péter Catholic University and am currently working on my PhD thesis.

>

What is the topic of your PhD and why did you decide to focus on this problem? How are you approaching it and what do you hope to achieve with it once it is completed, both in terms of scientific results and to your research community? What are you currently busy with?

<

The topic of my PhD is coreference resolution, which is widely researched within the scope of computational linguistics. However, I assume that I could show something new because my approach differs slightly from the classical view of computational linguistics. The reason for that is that my way to computational linguistics has led through classic humanities and theoretical linguistics, therefore I investigate this topic rather as a theoretician but I keep in mind the applicability as well.

Currently I am building a coreference corpus which – beyond the usual analysis layers such as tokenization, part-of-speech tagging, morphological analysis and dependency parsing – will contain anaphoric and coreference relationships. In the example ‘*I called my mother. She was really tired.*’ the personal pronoun ‘she’ refers back to its antecedent ‘my mother’ and this relationship is called anaphora. In contrast, coreference occurs when two expressions have the same referent and there are numerous forms of this relationship (e.g. repetition, name variants, synonymy, part-whole relationship, etc.). In the example ‘*I bought a bicycle. Tomorrow I will ride home my new bike.*’ the base of the coreference relationship between ‘bicycle’ and ‘bike’ is synonymity.

Anaphora and coreference show similar behaviour across languages. However, in contrast with English, Hungarian is a pro-drop language, which means that some pronouns (namely the personal and possessive pronouns) can be dropped from the sentence following fairly subtle rules. In these cases, the person and number of the subject and the object can be calculated from the inflection of the finite verb, and the person and number of the possessor are calculable from the inflection of the possessed, therefore the use of zero pronouns can be handled in a simple rule-based manner. As a zero pronoun can also refer back to its antecedent, it needs to be indicated in the coreference corpus. I have created an application that inserts the dropped pronouns into the texts, therefore these pronouns can also play a role in anaphora resolution. The corpus could serve as a resource for further research on this topic, be it answering theoretical questions or a technical application for a certain purpose.

Building a corpus of gold standard quality is definitely complicated and time-consuming. But still, the process of corpus building allows one to study the object of anaphora and coreference very meticulously. The feedback of my annotators also gives lessons to be learned. Therefore, together with the corpus, I increase my own knowledge about the phenomena. At the end of the pilot phase I am going to be in possession of the know-how that allows further enlargement of the resource.

>

How did you get involved with HUN-CLARIN and what is your experience with it?

<

My department has multiple connections with HUN-CLARIN. Firstly, the Old Hungarian Corpus (http://oldhungariancorpus.nytud.hu/) was produced in my institute. Initially, I was involved in this project as an annotator, I manually corrected the output of the optical character recognition on Old Hungarian texts. Later, to speed up the work, I developed a small script for helping manual normalization (standardization of old or non-standard texts). It turned out that manual work could be considerably cut down with the help of this pre-normalization tool.

Secondly, I am involved in the e-magyar project, a text processing pipeline for Hungarian, which is also connected to HUN-CLARIN. Last year I developed two small but useful modules for e-magyar, both of which are responsible for conversion between certain formats. One of them converts from the e-magyar tagset to an international standard part-of-speech tagset of Universal Dependencies (UD). The converter is needed for intermodular communication inside the pipeline, but could also serve as a useful output formalism due to the prevailing nature of UD. The other converter is applied between the internal format of e-magyar and the CoNLL-U format, a widely used international standard. The conversion between these two formats allows further work, annotation or visualization of the output with other tools related to the ConLL-U format. Both of the converters were needed for my own purposes in my corpus building project, but soon it turned out that the covered formalisms could be useful for other users as well, and therefore the converters have now been integrated in the e-magyar framework.

>

In addition to contributing to the development of e-magyar, you also have extensive experience in using it in practice. Could you briefly describe this dual role of yours and the advantages it brings?

<

I have a double relationship with e-magyar as I am an everyday user of it and a member of the developer team as well. This duality brings benefits: on the one hand, my needs are fulfilled thanks to the work of my colleagues, and on the other hand, my everyday experience with e-magyar serves as a useful feedback, which is important for maintenance and further development of e-magyar.

I use e-magyar principally in my corpus building project. Initially, the selected texts are analysed with the tokenizer, morphological analyser and part-of-speech tagger modules of e-magyar. Then, the output of e-magyar must be corrected manually, because the quality of the other annotation layers can be strongly influenced by the initial step. Next, the texts with the corrected annotation layers are further analysed by a sentence parser module of e-magyar, which produces the dependency trees of the sentences. This layer needs manual correction as well. At this point of the workflow, the texts with the corrected annotations are accessible for further, higher-level analysis, such as anaphora resolution.

I am working on three applications in connection with my PhD. The first one is responsible for inserting zero pronouns, the second resolves anaphora and the third resolves coreference. Indeed, the output of these applications also needs manual correction, but finally, besides a high-quality gold standard corpus, I obtain valuable observations of the quality of my applications. I hope that these three applications can also be added to the e-magyar chain as modules in the future.

>

You have recently also been involved in the development of Normo, a tool for the normalization of historical Hungarian. How is historical Hungarian different from contemporary Hungarian and why is such a tool needed? How does it work and who it is intended for? Has it been used on any text collection that is important for Hungarian humanities researchers?

<

Normalizing old texts is an important step in the workflow, because of the heterogeneity of the old orthographic system applied in historical texts. Normalization makes the text readable for humans and also for computers. There are multiple approaches to normalization – our project aimed to preserve the structures of the old language variety making them investigable for historical linguists, and therefore the task of normalization here mainly means the standardization of the spelling (thus covering the differences between the Middle and Modern Hungarian alphabet).

Since manual normalization is time-consuming and requires highly skilled and delicate work, application of automatic methods can help a lot. According to our measurements and the feedback from our annotators, Normo, our pre-normalization tool, eases and shortens the manual normalization work.

Normo consists of two main modules. the first one is a memory-based module with a relatively small dictionary of the most frequent words in the New Testament and their modern equivalents. Based on this dictionary, the most frequent words can simply be replaced with their modern forms. The second one is a rule-based module which works with manually defined rewrite rules. These rules come from two sources: some of them were defined on the basis of known changes in the history of Hungarian, others were defined through corpus-based observations. While the character-level rules are applied inside a word (e.g. replacement rules for handling characters that are not used in Modern Hungarian), the so-called token-level rules operate across word boundaries for splitting or joining words according to the rules of the modern orthography. Normo has been used in the project of building the Old Hungarian Corpus and has been applied to our five Middle Hungarian Bible translations.

>

What are your plans and dreams for the future?

<

My biggest future plan is to work further on my coreference corpus and to make it available for others. With this it will be all set for seeking answers to other exciting questions. I also have to write up my dissertation. Apart from the work on my PhD I have recently been working on some other topics a lot. For instance, I became interested in morphological tagsets. I assume that I could exploit my theoretical–computational hybrid attitude in this field as well. Lastly, I have some favourite topics which I have already been working on (e.g. authorship attribution), I would like to work further on these topics later.

>

How can research infrastructures such as HUN-CLARIN best serve early-stage researchers and how can a new generation of researchers best contribute to the research infrastructure?

<

Recently, for example, I have attended a CLARIN workshop on NLP tools for historical data, which was a great opportunity for me. On the one hand, the event gave me a chance to get to know other researchers of a specific field. On the other, as I think it is essential for beginners to gain self-confidence among their colleagues, which comes gradually through presenting your research often, not to mention the fluent use of English. Additionally, CLAIRN conferences and workshops serve as a good platform to share new ideas with colleagues who have more experience and get useful feedback. The world of conferences, workshops and networking is of course only one aspect of the CLARIN infrastructure's benefits. However, according to my recent experiences, it is one really worth mentioning.

>

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume II
(publication number: CLARIN-CE-2019-1537, November 2019).

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2019-1537
November 2019

ISBN
9789082990911

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands

www.clarin.eu

CLARIN
Common Language Resources and
Technology Infrastructure

CLARIN 2019