Building a Manually Annotated Hungarian Coreference Corpus
Workflow and Tools

Noémi Vadász

CRAC Workshop 2022

Roadmap

1. main information about KorKor
2. the steps of the workflow
3. further questions

Main Information about KorKor

KorKor

- multiple linguistic annotation layers: morphological disambugiation in two tagsets, dependency parsing, zero verbs and dropped pronouns, anaphora and coreference
- all annotations are corrected manually by linguists
- two file formats
- all data, scripts and guidelines are available here: `https://github.com/vadno/korkor_pilot`
- CC-BY-4.0. license

	documents	tokens (`conllup`)	tokens (`xtsv`)
huwiki	62	16,739	18,262
globv	32	7,760	8,799
TOTAL	94	24,499	26,581

The Steps of the Workflow

The Workflow

1. text collection
2. `emtsv` process (`emToken`, `emMorph`, and `emTag` modules)
3. format conversion
4. manual check (Google Spreadsheets)
5. format conversion
6. `emtsv` process (`emDep` module)
7. format conversion (`emCoNLL` module)
8. manual check (WebAnno)
9. manual insertion of zero substantives and ellipted verbs (plain text editor)
10. zero pronoun insertion (`emZero` module)
11. pronominal anaphora resolution
12. manual check and coreference annotation (Google Spreadsheets)
13. format conversion

Preprocessing, Tokenization and POS-tagging

- Wikipedia and news articles (from OPUS Corpus)
- texts of manageable sizes without truncation
  the length of the documents are between 5 and 27 sentences
- spelling errors were corrected
- `emtsv`
  file format: `xtsv`, a TSV file with a header, one token per line, empty lines separate the sentences
  4 columns for the linguistic annotations: token, all possible morphological tags and lemmata, disambiguated lemma and disambiguated morphological tag

Manual Correction

- tokenization, disambiguated lemmata and morphological tags were corrected manually
- seven linguists
- Google Spreadsheets documents with highlighting and conditional formatting
- for each token the annotator had to check the disambiguated tag and lemma
- if it was not correct, he/she could choose from the other alternatives by the morphological analyzer
- if none of them were correct, he/she could set it manually
- tokenization errors were corrected by commands in certain cells of the spreadsheet
- a postprocessing script carried out the correcting commands (e.g. line deletion, line insertion, joining or splitting tokens etc.)

Morphological Tagsets: emMorph and UD

emMorph

- the output of emMorph morphological analyzer
- full morphosyntactic analysis (lemma, POS, features)
  adtad ('you gave it')
  lemma: ad
  POS and features: [/V][Pst.Def.2Sg]

UD

- Universal Dependencies (UD) tagset, v2 version
- the input of emDep, the dependency analyzer in emtsv
- less detailed as emMorph, it can be converted from emMorph with a module of emtsv
- universal and language specific linearized feature–value pairs *adtad* ('you gave it')
  POS: VERB
  features: Definite=Def|Mood=Ind|Number=Sing|
  Person=2|Tense=Past|VerbForm=Fin|Voice=Act

Dependency Parsing, Manual Correction and Zero Substantives

- dependency analysis by `emDep`
- manual correction in WebAnno by three linguists
- the output of `emDep` was converted to CoNLL-U for WebAnno by a module of `emtsv`
- some zero tokens were inserted into the dependency trees: zero substantives and ellipted verbs
- zero substantives also have a subject and ellipted verbs also have an object or other arguments
- 419 zero substantives and 22 ellipted verbs were inserted into the corpus

Inserting Dropped Pronouns

- Hungarian is a pro-drop language, the subject, the object and the possessor can be dropped in some cases
- a rule-based script inserted the dropped pronouns
- the rules work on the preceding annotation layers (lemma, morphological tag and dependency analysis) in the following cases:
  - subject, if a verb does not have a subject in the dependency tree;
  - object, if a transitive verb does not have an object in the dependency tree;
  - possessor, if a possessum does not have a possessor in the dependency tree;
  - subject for an inflected or a non-inflected infinitive in the dependency tree.
- this step created extra branches in the dependency trees
- the morphological features of the inserted pronouns (person and number) are calculated from the verb or the possessum
- 867 zero subjects, 101 zero objects and 379 zero possessors were inserted

Inserting Pronominal Anaphora

- pronominal anaphora relations are inserted by a rule-based script
- the script searches for personal pronouns
- a set of rules operate on the POS-tag, the morphological features and syntactic information
- it searches for an antecedent
- e.g. if the subject of a verb is covert and the inflection of the verb is identical to the verb of the previous clause, the antecedent of the subject is the subject of the verb in the previous clause

Manual Correction and Coreference Annotation

- 4 linguists checked and corrected the insertion of dropped pronouns and pronominal anaphora and they annotated coreference relations in this phase
- Google Spreadsheets with conditional formatting
- anaphora types: personal (**prs**), demonstrative (**dem**), reciprocal (**recip**), reflexive (**refl**), relative (**rel**), possessive (**poss**)

Generic Subject

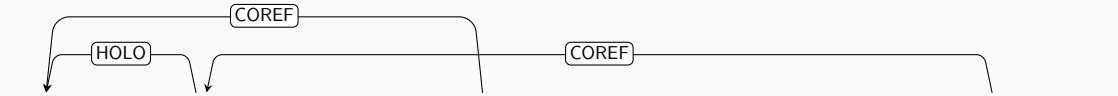
a Kínai Kommunista Párt egyik volt vezetője, akit hazaárulás miatt **elítéltek**
*one of the ex-leaders of the Communist Party of China, who was **convicted** for treason it was first **mentioned** in 1883 as an area donated to the Orthodox community*

Speaker and Addressee

A születésnapi ajándékoknak is nagyon **örülünk**, ha **szeretnéd** támogatni a munkánkat, **küldj nekünk** adományt, vagy **vegyél** egyet az NSA-s karácsonyi **üdvözlőlapjaink** közül, amelyet a Creative Time-nál dolgozó **barátaink** terveztek.
***We**'re also very happy for birthday gifts, if **you** want to support **our** work, send us a donation, or buy one of our NSA Christmas cards designed by **our** friends at Creative Time.*

Coreference and Branching Coreference Chains

- coreference types: the two elements have identical reference (**coref**), a part-whole relation holds between the two entities (**holo**)
- branching chains: while in a coreference relation both participants are overt, the antecedent of a pronoun can be either a dropped pronoun or an overt phrase, therefore anaphoric and coreference relations make up a tangled net with branches, instead of a simple chain



Converting to CoNLL-U Plus

`xtsv`

- it is the format of `emtsv`, a Hungarian text processing pipeline
- `tsv` with header
- one token per line
- empty line separates the sentences
- the annotations are in the columns, which are described in the header

`conllup`

- `conllup` files were converted from `xtsv`
- CoNLL-U Plus file format of Universal Dependencies
- `tsv`
- one token per line
- empty line separates the sentences
- the annotations are in the columns, which are described in the first comment line

Differences between the two formats

- zero elements are separate tokens in `xtsv`, but not in `conllup`
- zero substantials and ellipted verbs are not annotated in `conllup`
- dropped pronouns are marked in a separate column of `conllup` in the line of the verb or the possessor
- dropped pronouns are left out from the coreference chains in `conllup`

→ there are differences between the dependency trees and anaphoric chains between the two formats

`xtsv`: suitable for examining the nature of anaphora from the linguistic point of view

`conllup`: it is more applicable as a training or a test dataset, therefore is can form a base of a higher level information retrieval task

Further Questions

- The state of the referent changes: what kind of relationship exists between a human and his/her dead body?
  Three months have passed since the murder of ***the journalist couple****, Sagar Sarwar and his wife.* ***The bodies*** *are already exhumated to repeat the autopsy.*
- split antecedents
  Papyrus *is brave and saves* ***Thèti-Chèri****.* ***The two friends found each other*** *got a mission from the gods to guard the pharaoh.*
  Papyrus and Thèti-Chèri *lost each other.* ***The two friends found each other*** *got a mission from the gods to guard the pharaoh.*

Thank you for your attention!