CG to Dependency Parse

- Reusing Grammatical Resources for New Languages
- Estonian Dependency Treebank: from Constraint Grammar Tagset to Universal Dependencies

Reusing Grammatical Resources for New Languages

Lene Antonsen, Trond Trosterud, Linda Wiechetek

Takeaway

Machine-readable grammars can be more easily applied to new langauges if they are working with higher levels of analysis. Working with morphophonology, the grammatical differences between languages preclude the reuse of analyses.

> “ We argue that portability here takes the form of reusing smaller modules of the grammar

Hopefully the paper expands on that because that statement doesn't make any sense

Languages

-

North Lule and South Sami
    - Uralic language
    - Not very agglutinative
- Faroese
    - Germanic language
    - Four case system
- Greenlandic
    - Eskimo-Aleut language
    - Polysynthetic

Technical background

- Using existing resources developed by the University of Tromso.
    - Morphological analyzers
    - Constraint Grammar parsers

Reusing grammar

- Blick (2006) argues for using bootstrapping techniques to reuse grammar instead of appealing to statistical systems. *This fell by the wayside, everyone uses statistical methods now*

The bottom of the analysis

- The level of analysis that is close to the language substance cannot be directly used
- > “

> Even though different languages do not have the eact same morphological processes, they may have the same process *types*

- Rules are written in a modular fashion so they can easily be adapted to new languages
  - For example, consonant gradiation processes are very common, the particulars of the rule may need to change but the module design helps guide the changes that need to be made.

Disambiguation

Mapping of syntactic tags

- Large number of tags needed due to the free word order of Sami languages
  - For example, four different subject tags needed specifying whether the verb is finite, whether elipsis of verb has occured, whether the finite verb is to the left or to the right etc.
-

The top of the analysis

This is the part that's relevant to me

- Using a constraint grammar module
-
  > “ Syntactic tags for verbs are substituted by other tags (according to clause-type) in order to make it easier to annotate dependency across clauses

- Descibes difficulties finding the "head" of the sentence (think they mean root), when dealing with ellipses. This is definitely an issue as well in UD

> “ Still the analyzer retains very good accuracy for the dependency analysis: 0.99

- This is for Sami
- Table 5 say this is actually f-score?
- How is this scored? Are they scoring the flat descriptors in the visl format (e.g. #5->0)
- Use pairs of substitution and setparent rules

Bootstrapping

- Go through small modifications to the rules to consider Faroese specific phenomenon.
- Show the specific increases in performance with each new difference that is considered (e.g. when substituting the Relative pronouns that begin subordinate clauses in Sami with the CS that begins relative clauses in Faroese, the accuracy goes up to 96)

Estonian Dependency Treebank: from Constraint Grammar Tagset to Universal Dependencies

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen

Central topic

Dependency treebank in Universal Dependencies formalism adapted from an existing dependency treebank for Estonian. this adaptation was doen semi-automatically using a Constraint Grammar transfer rule system.

Methodology

Structure of annotations

The Estonian Dependency Treebank (DT) is annotated in Constraint Grammar style. There are

three layers:

- morphological
- surface syntactic
- dependency

This is an example word tag in a larger sentence (for more information see Figure 1 in the paper).

```
"<lamnast>"
   "lamnas" Lt S com sg part @<Q #6->5
```

> The used set of syntactic relations derives from Constraint Grammar, but the definitions of syntactic relations...are based on an academic description of Estonian grammar

Differences between UD and EDT annotation

Both EDT and UD adopt dependency grammar-based annotation guidelines. However, different syntactic relations are used and some phenomena are analyzed differently.

POS tags

No DET tag in estonian UDT, smilar decision made for The Fininnish UDT. PART not used because these things are currently tagged as adverbs or pronouns and it would require manual effort to retag them.

No discussion of annotation of morphological features.

Ditransitives are not used as there are no grammatical descriptions of Estonian that describes

ditransitives in Estonian.

EDT distinguishes between finite and non-finite (*subordinate*) clauses with finite clauses not indicating the syntactic relation between the head of the finite clause and the main clause *what are they doing here then? This is very unclear in this paper, maybe I need to read the paper for the EDT in order to make sense of this*.

EDT annotated modals and other auxiliaries as multi-word predicates. Many of these are set up as complementary clauses with ccomp and xcomp in UD instead.

Primacy of content words in UD causes a large number of changes. EDT did a lot of relations between functional words. For example, nouns in a prepositional phrase were dependents of the preposition, while the preposition was dependent on the larger context. In UD, this has to be changed because dependency relations need to be between content words.

Conversion procedure

- Rearrange subtrees, find connections between UD and EDT *I thought this was manual exploration of differences first, but it does appear this is the actual tree rewriting*
  - Using Vislcg3 *like I intend to*
- Convert from CG3 format (default in ED%) to CONLL-U, convert pos tags, morphological features using simple mapping.
- Formal checks to verify there is one and only one root, verify valid dat, all fields filled in.

Findings

Estimation of conversion quality:

- Used MaltEval
- UAS of 96.3
- LAS of 98.4
- annotation of punctuation marks was an issue.

- ccomp is the most error prone dependency relation at 64.3%

UD's emphasis on dependencies between content words results in projectivity (often). Where EDT was non-projective, the UD version is projective.

Follow up readings