A Universal Dependencies compatible Syntax-Treebank for Welsh

Orange Labs, TGI / Data & IA / DESKIÑ

Johannes Heinecke
15<sup>th</sup> Gyngres Astudiaethau Celtaidd Ryngwladol (ICCS) 2019
Table of Contents

1. Treebanks

2. Universal Dependencies
   - Origin, goals
   - UD Treebanks
   - Morphological and syntactical categories
   - Format
   - Flat trees

3. The Welsh Treebank
   - Annotation process
   - Tool
   - Evaluation
   - Welsh Dependency Syntax

References
Treebanks

Corpus of sentences as syntax trees

- **Phrase Structure Grammar**: Penn Treebank (Marcus, Kim, et al. 1994)

```
S
└── NP-SBJ-I
    └── Chris

S
└── VP
    ├── wants
    └── NP
        └── the ball

S
└── VP
    ├── NP
    │   └── to
    └── nsubj
```

- **Dependency grammar**: UD

```
S
├── root
├── nsubj
└── xcomp

S
├── want
├── to
└── nsubj
```

- **Semantic Treebanks**
  - Abstract Meaning Representation (AMR)
  - Universal Conceptual Cognitive Annotation (UCCA)
  - FrameNet
  - PropBank
  - DeepBank (HPSG)

- **Usually manually annotated and validated by linguists**
framework for cross-linguistically consistent grammatical annotation

- cross-linguistically **consistent treebank annotation** for many languages
- facilitating **multilingual** parser development, **cross-lingual** learning, and parsing research from a language typology perspective
- provide a **universal inventory of categories and guidelines** to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary.
- everybody can contribute (editing, validating, correcting, ...)
UD Treebanks (version 2.4 as of May 15, 2019)

- currently 146 treebanks in 83 languages
- including small and « low resource » languages (Romansh, Uyghur, Warlpiri) and ancient languages (Akkadian, Ancient Greek, Latin, Gothic, Sanskrit, Old Church Slavonic)
- different text genres (news, wikipedia, literature, religious texts)
- some sign languages, and corpora of (transcribed) spoken language
UD Treebanks III

Languages

- most of the data with free licences (like Creative Commons)
  - Afrikaans, Akkadian, Amharic, Ancient Greek, Arabic, Armenian, Assyrian, Bambara, Basque, Belarusian, Breton, Bulgarian, Buryat, Cantonese, Catalan, Chinese, Classical Chinese, Coptic, Croatian, Czech, Danish, Dutch, English, Erzya, Estonian, Faroese, Finnish, French, Galician, German, Gothic, Greek, Hebrew, Hindi, Hindi English, Hungarian, Indonesian, Irish, Italian, Japanese, Karelian, Kazakh, Komi Zyrian, Korean, Kurmanji, Latin, Latvian, Lithuanian, Maltese, Marathi, Mbya Guarani, Naija, North Sami, Norwegian, Old Church Slavonic, Old French, Old Russian, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovak, Slovenian, Spanish, Swedish, Swedish Sign Language, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Upper Sorbian, Urdu, Uyghur, Vietnamese, Warlpiri, Welsh, Wolof, Yoruba.

- 20 language families/language types
  - Creole, Code switching, Sign Language

- Still missing: Amerindian languages, Caucasian languages (NW, NE, Kartvelian), Bantu languages, Polynesian languages
Size of the existing treebanks

Version 2.4

- treebanks varying in size from 55 sentences up to 206,794 sentences or 3,800,000 tokens (German), with an average of 8,456 sentences (150,827 tokens) and a median of 2,802 (43,754)

- Celtic languages:
  - Irish: 1,020 sentences, including questions (Lynn and Foster 2016)
    - 24,000 tokens (average: 23.5 words per sentence, median length: 19, shortest sentence: 1, longest sentence: 311)
  - Breton: 888 sentences (Tyers and Ravishankar 2018)
    - 10,348 tokens (average: 11.7, median length: 10, shortest sentence: 2, longest sentence: 40)
  - Welsh: 601 sentences,
    - 10,756 tokens (average: 17.9, median length: 16, shortest sentence: 4, longest sentence: 59)
UD Treebanks II

UD Welsh
Johannes Heinecke
ICCS 2019

Treebanks
Universal Dependencies
Origin, goals
UD Treebanks
Morphological and syntactical categories
Format
Flat trees
The Welsh Treebank
References

UD Treebanks II
UD Treebanks II

Treebanks
Universal Dependencies
Origin, goals
UD Treebanks
Morphological and syntactical categories
Format
Flat trees
The Welsh Treebank
References
UD Treebanks II

Treebanks
Universal Dependencies
Origin, goals
UD Treebanks
Morphological and syntactical categories
Format
Flat trees
The Welsh Treebank
References
Tokenisation

- following the orthographic standard of each language (Chinese: separated lemmas)
- multi word entities *après-midi* ("without" internal syntax)
- punctuation as individual tokens
- some contracted forms are decomposed (e.g. French *du* > *de le*, Welsh *yr/nid* + inflected form of *bod*, Arabic prepositions/conjunctions, possessives or object clitics)
### 17 universal part-of-speech (UPOS) tags (not all are present in every language)

<table>
<thead>
<tr>
<th>Open class words</th>
<th>Closed class words</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>ADP</td>
<td>PUNCT</td>
</tr>
<tr>
<td>ADV</td>
<td>AUX</td>
<td>SYM</td>
</tr>
<tr>
<td>INTJ</td>
<td>CCONJ</td>
<td></td>
</tr>
<tr>
<td>NOUN</td>
<td>DET</td>
<td>X</td>
</tr>
<tr>
<td>PROPN</td>
<td>NUM</td>
<td></td>
</tr>
<tr>
<td>VERB</td>
<td>PART</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRON</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCONJ</td>
<td></td>
</tr>
</tbody>
</table>

Additional language-specific XPOS tags (e.g. PRON: { dep, indep }) available.
## « Universal » categories III

### 23 morpho-syntactic features

<table>
<thead>
<tr>
<th>Lexical features</th>
<th>Inflectional features</th>
</tr>
</thead>
<tbody>
<tr>
<td>PronType</td>
<td>Gender</td>
</tr>
<tr>
<td>NumType</td>
<td>Animacy</td>
</tr>
<tr>
<td>Poss</td>
<td>NounClass</td>
</tr>
<tr>
<td>Reflex</td>
<td>Number</td>
</tr>
<tr>
<td>Foreign</td>
<td>Case</td>
</tr>
<tr>
<td>Abbr</td>
<td>Definite</td>
</tr>
<tr>
<td></td>
<td>Degree</td>
</tr>
</tbody>
</table>

- Not all features used in every language
- language specific features exist (e.g. for Welsh: Mutation, Tenses)
### « Universal » categories IV

#### 37 universal dependency relations (can be extended with language specific ones)

<table>
<thead>
<tr>
<th></th>
<th>Nominals</th>
<th>Clauses</th>
<th>Modifier words</th>
<th>Function Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core arguments</strong></td>
<td>nsubj</td>
<td>csubj</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>obj</td>
<td>ccomp</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>iobj</td>
<td>xcomp</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-core dependents</strong></td>
<td>obl</td>
<td>advcl</td>
<td>advmod</td>
<td>aux</td>
</tr>
<tr>
<td></td>
<td>vocative</td>
<td></td>
<td>discourse</td>
<td>cop</td>
</tr>
<tr>
<td></td>
<td>expl</td>
<td></td>
<td></td>
<td>mark</td>
</tr>
<tr>
<td></td>
<td>dislocated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nominal dependents</strong></td>
<td>nmod</td>
<td>acl</td>
<td>amod</td>
<td>det</td>
</tr>
<tr>
<td></td>
<td>appos</td>
<td></td>
<td></td>
<td>clf</td>
</tr>
<tr>
<td></td>
<td>nummod</td>
<td></td>
<td></td>
<td>case</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coordination</th>
<th>MWE</th>
<th>Loose</th>
<th>Special</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>conj</td>
<td>fixed</td>
<td>list</td>
<td>orphan</td>
<td>punct</td>
</tr>
<tr>
<td>cc</td>
<td>flat</td>
<td>parataxis</td>
<td>goeswith</td>
<td>root</td>
</tr>
<tr>
<td></td>
<td>compound</td>
<td></td>
<td>reparandum</td>
<td>dep</td>
</tr>
</tbody>
</table>
## Format

### Data format: CoNLL-U

Simple text files, a word per line, An empty line separates sentences.

<table>
<thead>
<tr>
<th>id</th>
<th>form</th>
<th>lemma</th>
<th>UPOS</th>
<th>XPOS</th>
<th>features</th>
<th>head</th>
<th>deprel</th>
<th>enh. rels</th>
<th>misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>Rydw</td>
<td>i'n</td>
<td>canu</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>1-2</td>
<td>Rydw</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>2</td>
<td>aux</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>1</td>
<td>yr</td>
<td>y</td>
<td>AUX</td>
<td>preverb</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>ydw</td>
<td>bod</td>
<td>VERB</td>
<td>verb</td>
<td>Ps=1</td>
<td>Num=Sg</td>
<td>Tense=Pres</td>
<td>0</td>
<td>root</td>
</tr>
<tr>
<td>3</td>
<td>i</td>
<td>i</td>
<td>PRON</td>
<td>indep</td>
<td>Ps=1</td>
<td>Num=Sg</td>
<td>2</td>
<td>nsubj</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>'n</td>
<td>yn</td>
<td>AUX</td>
<td>impf</td>
<td>_</td>
<td>5</td>
<td>aux</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>canu</td>
<td>canu</td>
<td>NOUN</td>
<td>verbnoun</td>
<td>VerbForm=Vnoun</td>
<td>2</td>
<td>xcomp</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>punct</td>
<td>_</td>
<td>2</td>
<td>punct</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>

Extended formats exist to add semantic and other information.
Critique

- too semantic
- too much English/Latin based: *nsubj*, *obj* of less use in ergative or active languages;
  “A nominal subject (*nsubj*) is a nominal which is the syntactic subject and the proto-agent of a clause... and this argument is the more agentive, the do-er, or the proto-agent of the clause.”
- auxiliaries depend on main verbs, which is a more semantic than syntactical view, since the subject agreement is still with the auxiliary
- prepositions dependants of nominals

- « flat trees », no dependants on auxiliaries: *Falch o gyhoeddi ein bod wedi bod yn llwyddiant gyda Cronfa Lawnsio*
Flat Trees? I

Yr ydyn ni 'n llywddiannus
PART AUX PRON PART ADJ

cop

advmod

nsubj

case:pred

root

comp

ccomp

PART AUX PRON PART ADJ
Flat Trees? I

15/33
Flat Trees? I

The Welsh Treebank

References
Usage

- **Computational Linguistics:**
  - train parsers (cf. parser evaluation tasks CoNLL-2017 (Hajič and Zeman 2017), ConLL-2018 (Zeman and Hajič 2018)) and annotate new texts (more or less correctly)

- **typology**

- **language comparison (syntax), c.f. Parallel Universal Dependencies (PUD) treebanks:**
  - 1,000 sentences, professionally translated into 18 languages (Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Thai, Turkish)
With population growth, new indigenous quarters were created.

Nüfus artşi ile yeni yerli mahalleler yaratılmıştır.

Gyda thwf poblogaeth, crewyd cwrtewau cynhenid newydd.
Annotation for Welsh: process and data I

Definitions

- annotation guide: Universal Dependencies, with extensions for Welsh:
  - language specific XPOS
    - type distinctions for named entities,
    - comparison degree for adjectives, verbnouns (= nouns),
    - subtypes for dependency relations having verbnouns as heads
  - Contracted verb formes decomposed: roedd → yr oedd

- origin of sentences in the treebank:
  - sentences arbitrarily chosen from Wikipedia (pages on Welsh places, objects etc. to avoid as maximally as possible pages contributed by non-native speakers)
  - BBC Cymru, Y Golwg
  - Proceedings of the Third National Assembly for Wales, (http://cymraeg.org.uk/kynulliad3/)
  - Welsh language pages edited by Welsh Universities
  - various Welsh bloggers
  - examples taken from grammars of Modern Welsh (King 1993; Thorne 1993)

- original orthography including dialectal variations: ydi/ydy/yw, yr ydwyf, rydw, dwi rydan ni/ryden ni, doedd/dooddan, sy/syd, hyna/hynaf
Annotation for Welsh: process and data II

Annotation Process

- Transformation of CEG (Cronfa Electroneg o Gymraeg, https://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en) into CoNLL-U format
- Annotation loop
  - Tokenize, lemmatise, POS tag new Welsh sentences
    - Manual annotation of dependency relations
    - Correction of the automatic generated lemmas, UPOS and XPOS
  - Validation programmes
    - Check morphological features with Eurfa (inflected verbs)
    - Add mutation to morphological features
    - Add morphological features for inflected prepositions and forms of bod
    - Check whether lemmas of inflected verbs, preposition and plural noun forms ≠ form
    - Check dependency relations have coherent heads and dependent POS (e.g. acl must have a nominal head)
    - Scripts provided by UD to validate formal aspects of CoNLL-U
  - Train UDpipe and test performance of tagger and parser (Labelled Attachment Score, LAS)
  - Add more sentences and restart again
Current Results

Evaluation metrics

- **POS Tagging/Lemmatisation:** Precision (correct tag / all tags)
  - POS tagging (UPOS): 89.2%
  - Lemmatisation: 93.5% (using *Eurfa*).

- **Dependency analysis:**
  - Labelled Attachment Score (LAS): percentage of words (token) with correct head and dependency relation; 63.9% (on predicted POS), 76.2% (on gold POS).
  - LAS variants exist:
    - Weighted LAS: every dependency relation has a weight ($1 \geq w > 0$). “Less” important relations (*aux*, *case*, *cc*, *clf*, *cop*, *det*, *mark*) are downweighted.
    - Content Word LAS (CLAS): like weighted LAS, but relations not linking content words get a weight of 0 (Nivre and Fang 2017).

  Score calculated in terms of:
  - Precision (sum of weight of correct relations / sum of weight of all predicted relations).
  - Recall (sum of weight of correct relations / sum of weight of all correct relations).
  - F-measure:

\[
F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 54.8\%/69.6\% (\text{gold})
\]

- 5-fold cross evaluation.
Annotation Examples I

**Verbnouns**

- semantically events/states, syntactically nouns (UPOS: NOUN, XPOS: verbnoun)
  - cf. UD Irish Treebank, Welsh and Irish unlike UD Breton.

![Diagram of Welsh and Irish verb structures]

- patient marked in the same way as possessives using dependent pronouns (cf. *dy dŷ [ti]*)
- different to inflected verb forms, were nominal direct objects undergo soft mutation, and independent pronouns are used for pronominal direct objects.
- cf. German:
  - inflected: *ich sehe den Hund*<sub>ACC</sub>
  - infinitive: *den Hund sehen*<sub>ACC</sub>
  - verbnoun: *das Sehen des Hundes*<sub>GEN</sub>
  - “I see the dog”
  - “to see the dog”
  - “the sight/seeing of the dog”
- annotation of tense-aspect-markers differs
Annotation Examples II

Nested periphrastics

Other tense-aspect markers are annotated like wedi
- *am, ar, heb, newydd, ...*

Subordinate (no finite Tense)
**Annotation Examples III**

**Verbnoun bod**

- dependent pronoun (possessive) expresses *Agent* not *Patient*
- with nominal predicates, *bod* is a copula dependent

Comment:
Example with other verb
Annotation Examples IV

Inflected prepositions

- not contracted forms (at least synchronically), since the pronoun can follow:

```
Dyma 'r bachgen y rhoddodd hi wobr iddo
ADV DET NOUN AUX VERB PRON NOUN ADP
```

```
Dyma 'r bachgen y rhoddodd hi wobr iddo fo
ADV DET NOUN AUX VERB PRON NOUN ADP PRON
```
Inflected prepositions

- not contracted forms (at least synchronically), since the pronoun can follow:

```
Dyma 'r bachgen y rhododd hi wobr iddo fo
ADV DET NOUN AUX VERB PRON NOUN ADP PRON
```

```
Dyma 'r bachgen y rhododd hi wobr iddo fo
ADV DET NOUN AUX VERB PRON NOUN ADP PRON
```
**Annotation Examples V**

**cael**

- no passive annotation (nsubj:pass)
- main verbs depends on *cael* with an “clausal complement” (ccomp) relation

```
...Sut fyddwch chi 'n cael eich talu...
```

```
Mi gafodd o ei ladd...
```

```
root xcomp
  nsubj aux
    nsubj aux
      aux nsubj
        aux

root ccomp
  nsubj
    aux
      nsubj
        aux
```

**References**
Annotation Examples VI

Contracted Forms

- y/ni + form of bod (roedd, doedd, does, rydw, ...): annotated as two tokens (but kept as one contracted word)

```

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>advmod</td>
<td>cop</td>
<td>nsubj</td>
<td>advmod</td>
</tr>
<tr>
<td>root</td>
<td></td>
<td></td>
<td>case:pred</td>
</tr>
<tr>
<td>nid</td>
<td>oedd</td>
<td>hi</td>
<td>ddim</td>
</tr>
<tr>
<td>yrdd</td>
<td></td>
<td>yn</td>
<td>sal</td>
</tr>
</tbody>
</table>
```

Doedd

ADV AUX PRON ADV PART ADJ
Coordinated Numbers

```
<table>
<thead>
<tr>
<th>Coordination</th>
<th>Welsh Dependency Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>“21st century” (lit. “first century on twenty”)</td>
<td>“55 cars” (lit. “15 cars on 40”)</td>
</tr>
<tr>
<td>“61 days” (lit. “1 day on 60”)</td>
<td>“81 boats” (lit. “two boat(sg) and four twenধes”).</td>
</tr>
<tr>
<td></td>
<td>[my dependency annotation]</td>
</tr>
</tbody>
</table>
```

References

Treebanks
Universal Dependencies
The Welsh Treebank
Annotation process Tool
Evaluation Welsh Dependency Syntax
Conclusion

So far ...

- First Welsh treebank, third Celtic language treebank in UD
- 10,000 tokens, allows a relatively robust POS tagging (89.2%), and parsing of simpler sentences
Conclusion

So far ...

- First Welsh treebank, third Celtic language treebank in UD
- 10,000 tokens, allows a relatively robust POS tagging (89.2%), and parsing of simpler sentences

Next steps

- needs to be expanded
- more annotators needed!
- adding enhanced dependencies
References I


References II