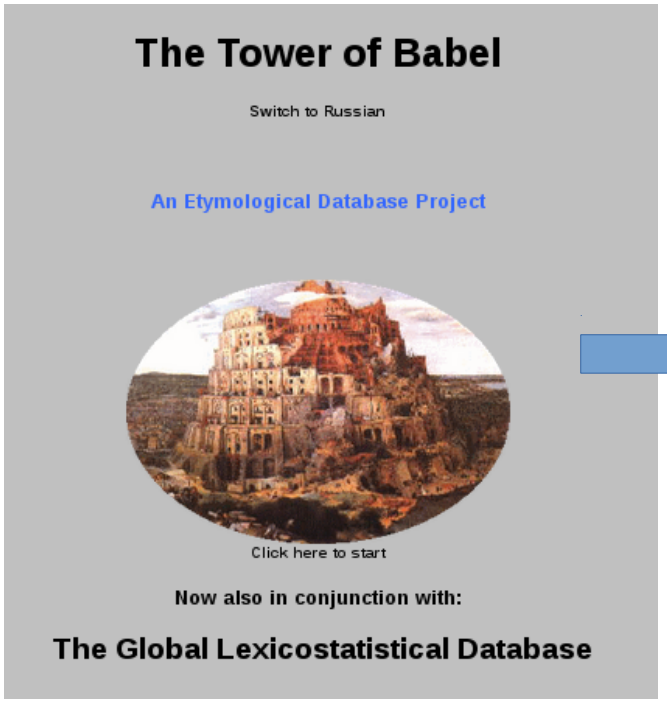
Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF

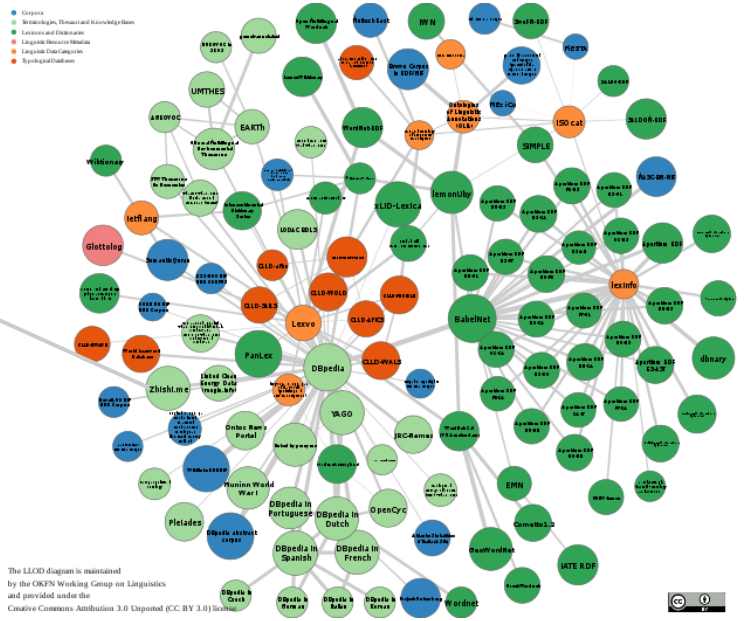
Frank Abromeit, Christian Chiarcos, **Christian Fäth**, Maxim Ionov

5th Workshop on Linked Data in Linguistics:
Managing, Building and Using Linked Language Resources
Portorož, Slovenia, 24th May 2016. Co-located with LREC 2016

Motivation



Lemon Ontolex



- Extending the LLOD Cloud with a large set of etymological resources
- Interoperability with a proprietary data format

Advantages of Linked Open Data

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- Reusability
  - Unique identifiers in the web of data (URI)
  - Standardized rich description formalisms like RDF and OWL
- Class / Type system
  - Easy to use with object oriented programming languages (e.g. for NLP)

- **lemon** (lexicon model for ontologies)
  - `http://www.w3.org/community/ontolex/wiki/Final_Model_Specification` (final version was released end 2015)
  - `https://www.w3.org/2016/05/ontolex/` (first official report)

The Tower of Babel

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

General information

- Web based project (`http://starling.rinet.ru`)
- Started by Sergei A. Starostin in 1998
- Historical and comparative linguistics
- Hosts over 50 etymological dictionaries

This talk's sample: Turkic etymological dictionary

- About 2200 entries
- Entries are derived from a reconstructed Proto-Turkic ancestor
- Cognate relationship of 29 languages

  Old Turkic, Karakhanid, Turkish, Tatar, Middle Turkic (Chagatai), Uzbek, Uighur, Sary-Yughur, Azeri, Turkmen, Oyrat, Khalaj, Khakassian, Chuvash, Yakut, Shor, Dolgan, Tuva, Tofalar, Kirghiz, Kazakh, Noghai, Bashkir, Balkar, Gagauz, Karaim, Karakalpak, Salar, Kumyk

The Tower of Babel – XML format

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- A downloaded dictionary can be converted to XML by using the *star4win* Windows application
  `http://starling.rinet.ru/download/star4win-2.4.2.exe`
- The structure of the XML is comprised of records for dictionary entries
- Dictionary data is encoded in XML as complex String values
- The XML structure is similar throughout all Starling dictionaries but encoding of dictionary data differs

Turkic etymological dictionary – XML format

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

```
<record id="6">
  <field name="NUMBER">6</field>
  <field name="PROTO">*Kuĺ</field>
  <field name="PRNUM">1157</field>
  <field name="MEANING">1 bird  2 duck</field>
  <field name="RUSMEAN">1 птица  2 утка</field>
  <field name="ATU">quš 1 (OUygh.)</field>
  <field name="KRH">quš 1 (MK, KB)</field>
  <field name="TRK">kuš 1</field>
  <field name="TAT">qoš 1</field>
  <field name="CHG">quš 1 (Sangl.); 'moth' (Abush.)</field>
  ...
  <field name="REFERENCE">VEWT 305, TMN 3, 547-548; EDT 670;
   ЭСТЯ 6, 180-182, Лексика 168, Stachowski 162. Chuv. хълat
  'hawk' &lt; Mong.
  </field>
</record>
```

Turkic etymological dictionary – XML format

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

```
<record id="6">
  <field name="NUMBER">6</field>
  <field name="PROTO">*Kuĺ</field>
  <field name="PRNUM">1157</field>
  <field name="MEANING">1 bird  2 duck</field>
  <field name="RUSMEAN">1 птица  2 утка</field>
  <field name="ATU">quš 1 (OUygh.)</field>
  <field name="KRH">quš 1 (MK, KB)</field>
  <field name="TRK">kuš 1</field>
  <field name="TAT">qoš 1</field>
  <field name="CHG">quš 1 (Sangl.); 'moth' (Abush.)</field>
  ...
  <field name="REFERENCE">VEWT 305, TMN 3, 547-548; EDT 670;
   ЭСТЯ 6, 180-182, Лексика 168, Stachowski 162. Chuv. хълат
  'hawk' &lt; Mong.
  </field>
</record>
```

Proto-Turkic form
→ Marked with asterisk
→ reconstructed

Turkic etymological dictionary – XML format

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

```
<record id="6">
  <field name="NUMBER">6</field>
  <field name="PROTO">*Kuĺ</field>
  <field name="PRNUM">1157</field>
  <field name="MEANING">1 bird  2 duck</field>
  <field name="RUSMEAN">1 птица  2 утка</field>
  <field name="ATU">quš 1 (OUygh.)</field>
  <field name="KRH">quš 1 (MK, KB)</field>
  <field name="TRK">kuš 1</field>
  <field name="TAT">qoš 1</field>
  <field name="CHG">quš 1 (Sangl.); 'moth' (Abush.)</field>
  ...
  <field name="REFERENCE">VEWT 305, TMN 3, 547-548; EDT 670;
   ЭСТЯ 6, 180-182, Лексика 168, Stachowski 162. Chuv. хълат
  'hawk' &lt; Mong.
  </field>
</record>
```

Meaning in Russian and English
- encoding **multiple** meanings
  - → 1 = bird
  - → 2 = duck

Cognates of up to 29 languages

Turkic etymological dictionary – XML format

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Cognate Fields

```
<field name="KRH">quš 1 (MK, KB)</field>
```

- For a cognate of a Proto-Turkish word the following information is stored
  - The proprietary language code (KRH for Middle Turkic)
  - At least one word form (quš)
  - (Optional) indexes (1) which refer to the word meaning as encoded in the MEANING/ RUSMEAN fields
  - (Optional) bibliographic references (MK, KB)
  - (Optional) gloss information to refine the word meaning as in the example below (recall meaning 1 = bird)

```
<field name="SHR">quš 1; 'hen'</field>
```

Turkic etymological dictionary – XML format

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

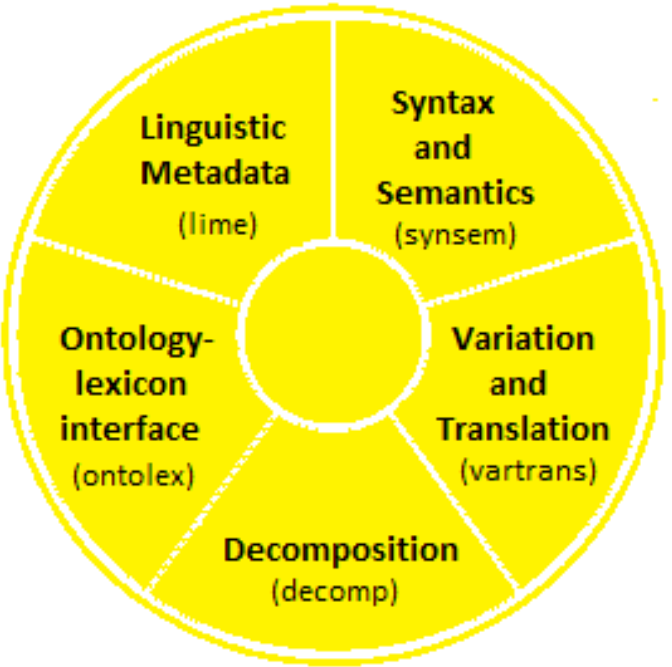
REFERENCE Field

- has bibliographic references for a Proto-Turkish word

```
<field name="REFERENCE">VEWT 305, TMN 3,
547-548; EDT 670; ЭСТЯ 6, 180-182, Лексика 168,
Stachowski 162. Chuv. хъlat 'hawk'<;
Mong.</field>
```

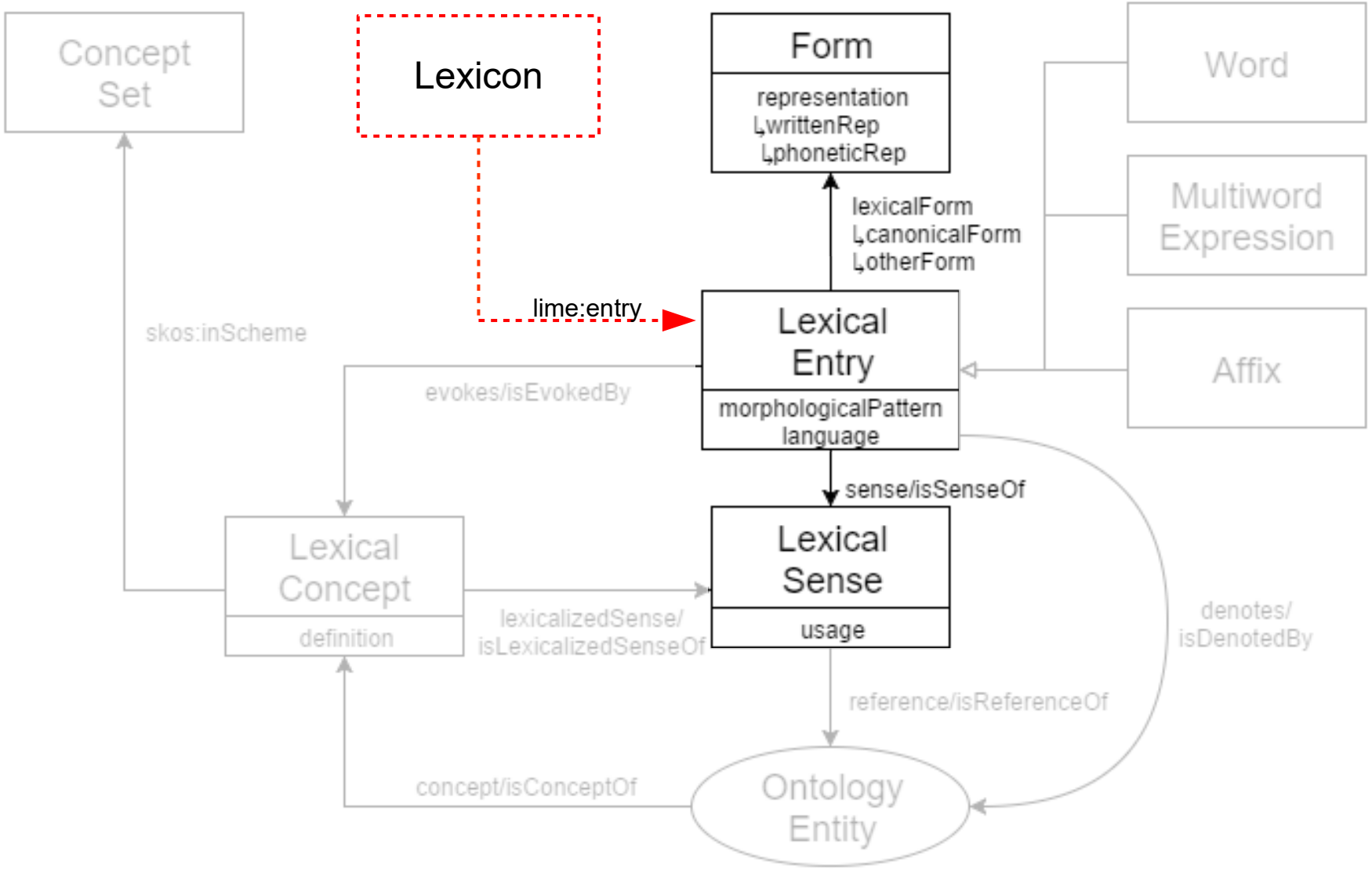
- Cited source given as abbreviation (VEWT)
- Location in cited source
- Gloss information e.g. to refine meaning (hawk)

Lemon modules



- For the Starling converter we use:
  - *ontolex* for lexical entries and lexical sense
  - *lime* for lexicon creation
  - *vartrans* for cognate relationships

Lemon / Ontolex core module



Converting the Turkic etymological dictionary

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Lemon lexicon

- For **each language** found in the dictionary a **separate lexicon** is created
- Lexicon entries are interlinked by means of RDF
- Language encoding:
  - **lime:language**: the original Starling encoding
  - **dct:language**: a manual mapping to lexvo.org

```
# Lexicon definition
star:lexicon_chg rdf:type       lime:Lexicon ;
                 dct:language   lexvo:chg ;
                 lime:language  "chg"
```

Converting the Turkic etymological dictionary

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Lemon lexical entry

- Words of a lexicon are represented in lemon as lexical entries
- An entry..
  - is created **for each proto- and cognate word**
  - can have several Forms and Senses
  - is added to the dictionary of its respective language

```
star:lexicon_chg/quš rdf:type lime:LexicalEntry ;
     ontolex:canonicalForm [ontolex:writtenRep "quš"] .

star:lexicon_chg lime:entry star:lexicon_chg/quš .
```

Converting the Turkic etymological dictionary

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Lemon lexical sense

- The Senses are only defined for Entries in the Proto-Turkic dictionary

```
star:lexicon_proto/*KuĺI/sense_1 rdf:type ontolex:LexicalSense ;
     skos:definition  "птица"@ru ;
     skos:definition  "bird"@en ;
     ...
```

- The Senses of their cognates reference the Proto-Turkic Senses

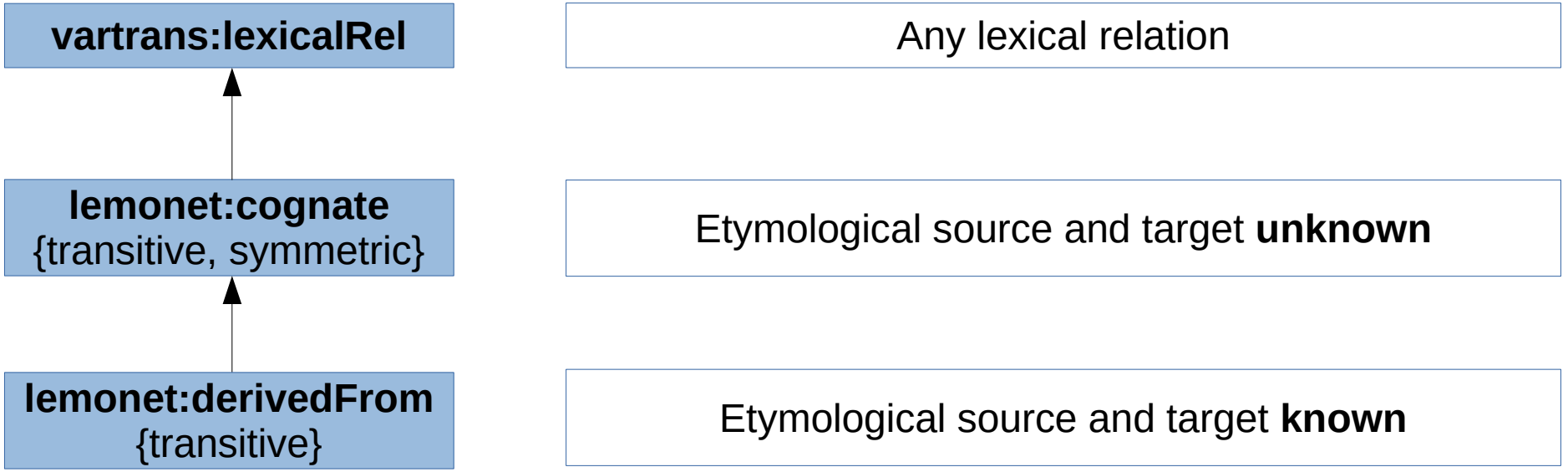
```
star:lexicon_chg/quš/sense rdf:type ontolex:LexicalSense ;
     ontolex:reference star:lexicon_proto/*KuĺI/sense_1 .
```

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Cognate modelling

- Namespace **lemonet** = 'lemon with etymological extensions' taken from Chiarcos, Sukhareva (2014)

```
star:lexicon_chg/quš
      lemonet:derivedFrom star:lexicon_proto/*Kuĺ
```



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Cognate gloss information

- Cognate fields may contain gloss information to further refine the meaning referenced by its index

```
<field name="CHG">quš 1 (Sangl.); 'moth'
            (Abush.)</field>
```

- These are included as rdfs:comment due to their complex, heterogenous nature

```
star:lexicon_chg/quš
     rdfs:comment "gloss : (Sangl.) and (Abush.)'moth'" .
```

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Bibliographic references

```
star:lexicon_proto/*Kuĺ
    dct:references star:lexicon_proto/*Kuĺ_/comment/VEWT .

star:lexicon_proto/*Kuĺ/comment/VEWT
    msh:cites            bib:VEWT ;
    rdfs:comment         "pages : 305" ;
    rdf:type             msh:Citation .

bib:VEWT  dct:date       "1969" ;
    talis:localityName   "Helsinki" ;
    dc:identifier        "VEWT" ;
    dct:isReferencedBy   "Altaic etymology, Turkic etymology,
                          Mongolian etymology" ;
    dc:title             "Versuch eines etymologisches
                          Wörterbuchs der Türksprachen." ;
    dc:creator           "Räsanen M." ;
    rdf:type             msh:Book .
```

Java converter

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- Converts Starling XML automatically to RDF
- Converter is applicable to all Starling etymological dictionaries
  - but parser has to be adjusted to match used encoding syntax and used XML field names
- freely available

`https://github.com/acoli-repo/starling-converter`

RDF-Conversion rates for Altaic dictionaries

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- Conversion results

	XML proto-forms	*XML cognates*	*Triples*
Turkic	82% (1651/2017)	77% (16726/21835)	145,981
Japanese	83% (1410/1705)	91% (5487/6009)	45,873
Korean	91% (1101/1206)	84% (1697/2025)	19,016
Mongolic	83% (1799/2173)	68% (7756/11328)	74,171
Tungus	81% (1963/2435)	83% (8260/9902)	66,549

- Converter was only optimized for Turkic
- Even without fine-tuning the parser, the results indicate relatively reliable extraction rates across different languages for both proto-form and cognate processing

Extensions

- Apply the converter to more Starling dictionaries
- Adjust parser to encoding variations of dictionary data
- Extract gloss meanings
- Link word senses to other LOD resources instead of encoding word meaning locally
  - For linking we consider lexical resources more appropriate than e.g. DBpedia or BabelNet as they cover a greater portion of the vocabulary
  - Use upcoming LOD version of the WordNet Interlingual Index (Bond et al., 2016, ILI)
  - Linking task is complicated by the sparsity of sense and gloss definitions in the Starling data

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Thank You!