Multi-dimensional Annotation and Alignment in an English-German Translation Corpus

Silvia Hansen-Schirra
Stella Neumann
Mihaela Vela

Universität des Saarlandes

Overview

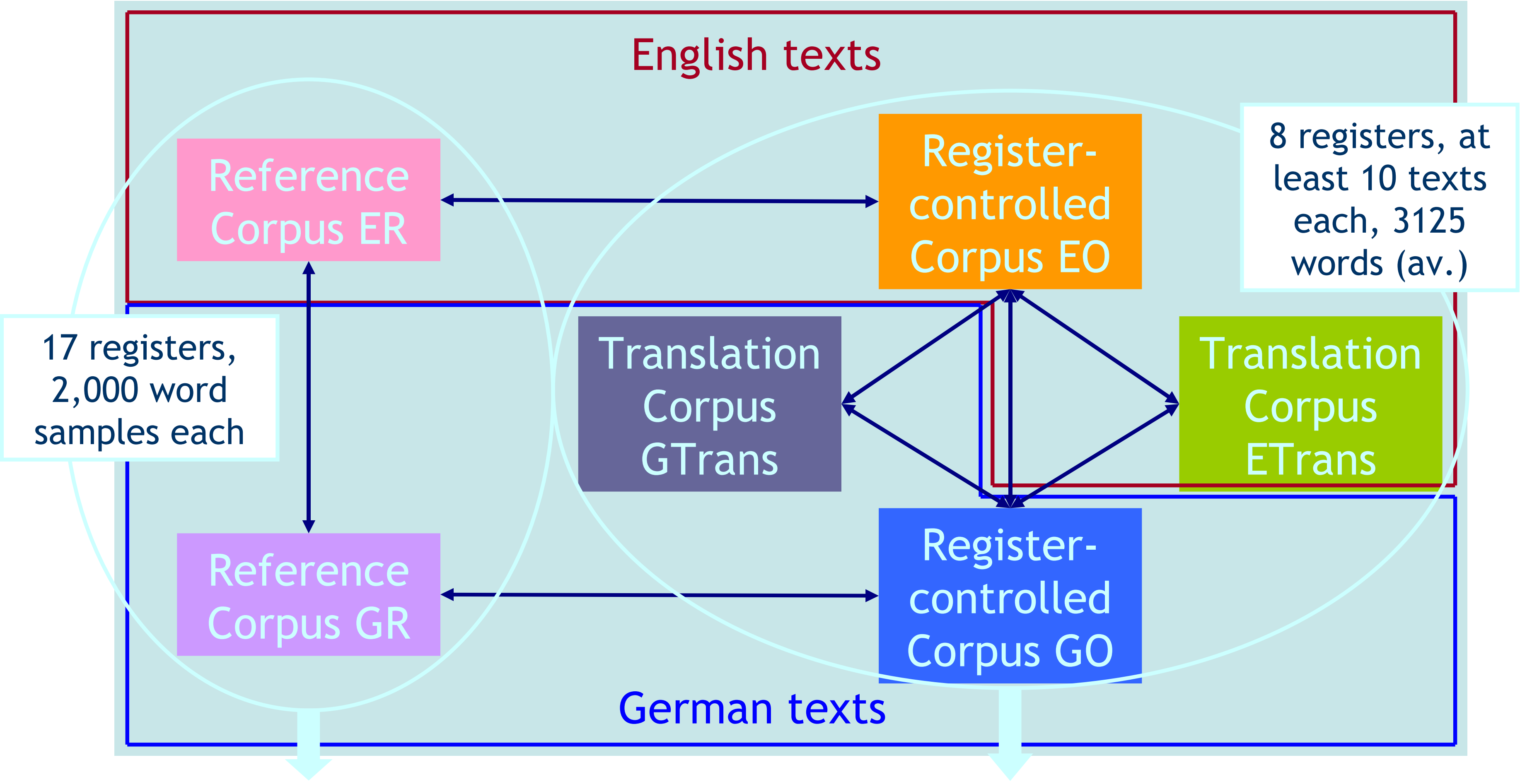
- CroCo
- Corpus Representation
- Linguistic Annotation
- Alignment
- Query
- Outlook

What is CroCo?

Corpus-based comparison of translations with originals in source AND target language

- Specific properties of translations: e.g. simplification, normalisation, explicitation
- Blum-Kulka 1986, Baker 1996, Olohan & Baker 2000

The CroCo Corpus


English texts
Reference Corpus ER
Register-controlled Corpus EO
8 registers, at least 10 texts each, 3125 words (av.)
17 registers, 2,000 word samples each
Translation Corpus GTrans
Translation Corpus ETrans
Reference Corpus GR
Register-controlled Corpus GO
German texts

68,000 words

1 million words

Corpus Representation

- File headers (Text Encoding Initiative)
  - Information about:
    - Author, publication, register information (text type)
    - Translator, translation process
- Text body (multi-layer stand-off XML)
    - Linguistic annotation and alignment

The Header

```
<teiHeader>
<fileDesc>
<filename>GO_FICTION_001.txt</filename>
<subcorpus>FICTION_GO</subcorpus>
<language>German</language>
<titleStmt>
<title>Mein Jahr als Mörder</title>
<author>Delius, Friedrich Christian</author>
</titleStmt>
<translation></translation>
<publicationStmt>
<publisher>Rowohlt Berlin Verlag</publisher>
<date>2004</date>
<distributor>http://www.litrix.de/mmo/priv/15719-WEB.pdf</distributor>
<availability>local</availability>
</publicationStmt>
<registerAnalysis>
…
</registerAnalysis>
…
</teiHeader>
```

Linguistic Annotation

- Morphology (MPro), part-of-speech (TnT), phrase structure (MPro), grammatical functions
- Representation format:
  - XML Annotation
  - Multi-layer: each annotation ➔ different layer
  - Stand-off annotation: annotation layers ➔ separate
  - Connection within a language by Xlink, Xpointer, xml:base attributes

XML Annotation

```
<document xmlns:xlink=
  http://www.w3.org/1999/xlink
  name="GO.tok.xml" xml:lang="de"
  docType="ori">
<header xlink:href="GO.header.xml"/>
 <tokens>
 <token id="t64" strg="Ich"/>
  <token id="t65" strg="spielte"/>
  <token id="t66" strg="viele"/>
  <token id="t67„
           strg="Möglichkeiten"/>
  <token id="t68" strg="durch"/>
  <token id="t69" strg=","/>
 </tokens>
</document>
```

```
<document xmlns:xlink=
  http://www.w3.org/1999/xlink
  name="GO.tag.xml">
 <tokens xml:base="GO.tok.xml">
  <token pos="pper"
           xlink:href="#t64"/>
  <token pos="vvfin"
   xlink:href="#t65"/>
  <token pos="pidat"
           xlink:href="#t66"/>
  <token pos="nn"
           xlink:href="#t67"/>
  <token pos="ptkvz"
           xlink:href="#68"/>
  <token pos="yc" xlink:href="#t69"/>
</tokens>
</document>
```

```
<document xmlns:xlink=
 http://www.w3.org/1999/xlink
  name="GO.chunk.xml">
 <chunks xml:base="GO.tok.xml">
  <chunk id="ch13">
   <tok xlink:href="#t66"/>
   <tok xlink:href="#t67"/>
  </chunk>
  <chunk id="ch14">
   <tok xlink:href="#t70"/>
  </chunk>
  <chunk id="ch15">
   <tok xlink:href="#t71"/>
  </chunk>
</chunks>
</document>
```

XML Annotation

```
<document xmlns:xlink=
 http://www.w3.org/1999/xlink
 name="GO.chunk.xml">
 <chunks xml:base="GO.tok.xml">
  <chunk id="ch13">
   <tok xlink:href="#t66"/>
   <tok xlink:href="#t67"/>
  </chunk>
  <chunk id="ch14">
   <tok xlink:href="#t70"/>
  </chunk>
  <chunk id="ch15">
   <tok xlink:href="#t71"/>
  </chunk>
 </chunks>
</document>
```

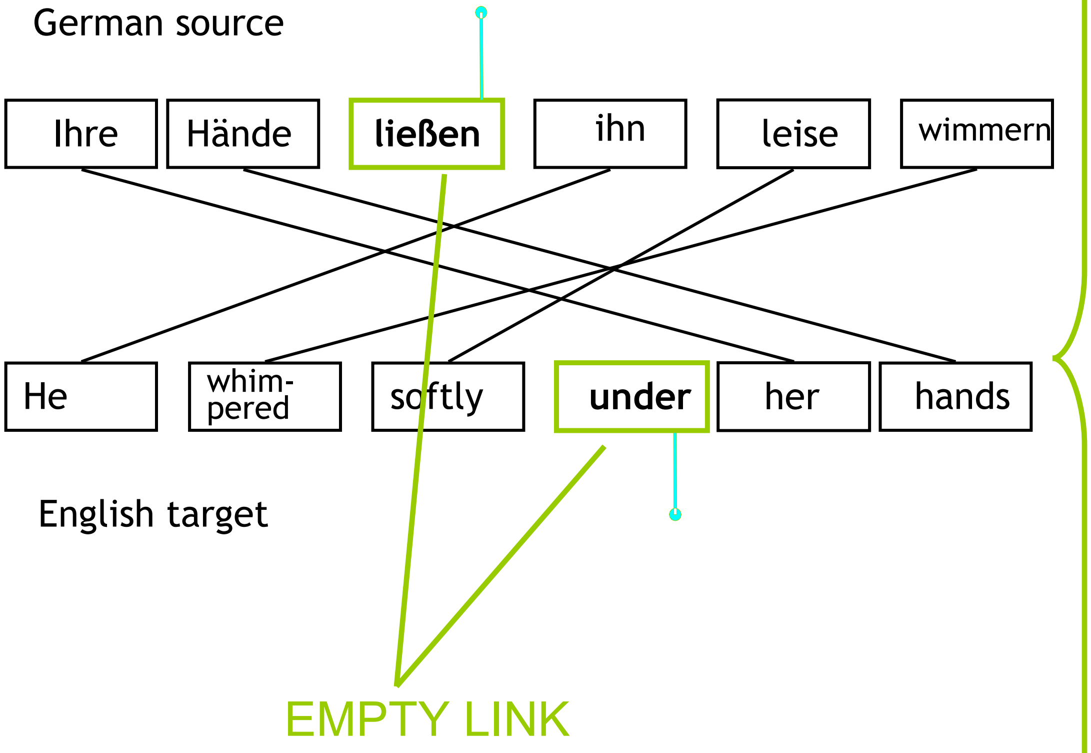
```
<document xmlns:xlink=
  http://www.w3.org/1999/xlink
  name="GO.ps.xml">
 <chunks xml:base="GO.chunk.xml">
  <chunk ps="NP"
     xlink:href="#ch13"/>
  <chunk ps="VPFIN"
   xlink:href="#ch14"/>
  <chunk ps="NP" xlink:href="#ch15"/>
  <chunk ps="NP" xlink:href="#ch16"/>
  <chunk ps="PP" xlink:href="#ch17"/>
  <chunk ps="NP" xlink:href="#ch18"/>
  <chunk ps="VPPRED"
   xlink:href="#ch19"/>
 </chunks>
</document>
```

```
<document  xmlns:xlink=
http://www.w3.org/1999/xlink
 name="GO.gf.xml">
 <chunks xml:base="GO.chunk.xml">
  chunk gf="DOBJ" xlink:href="#ch13"/>
  <chunk gf="FIN" link:href="#ch14"/>
  <chunk gf="IOBJ" xlink:href="#ch15"/>
  <chunk gf="DOBJ" xlink:href="#ch16"/>
  <chunk gf="ADV" xlink:href="#ch17"/>
  <chunk gf="PRED" xlink:href="#ch19"/>
 </chunks>
 </chunks>
</document>
```

Alignment

- Sentences (WinAlign, Trados), Clauses (MMAX II), Phrases (MMAX II), Words (GIZA++)
- Representation format:
    - XML Alignment
    - Multi-layer: each alignment ➔ different layer
    - Stand-off annotation: alignment layers ➔ separate
    - Connection between source and target language by Xlink and Xpointer attributes plus <translations> element

XML Token Alignment

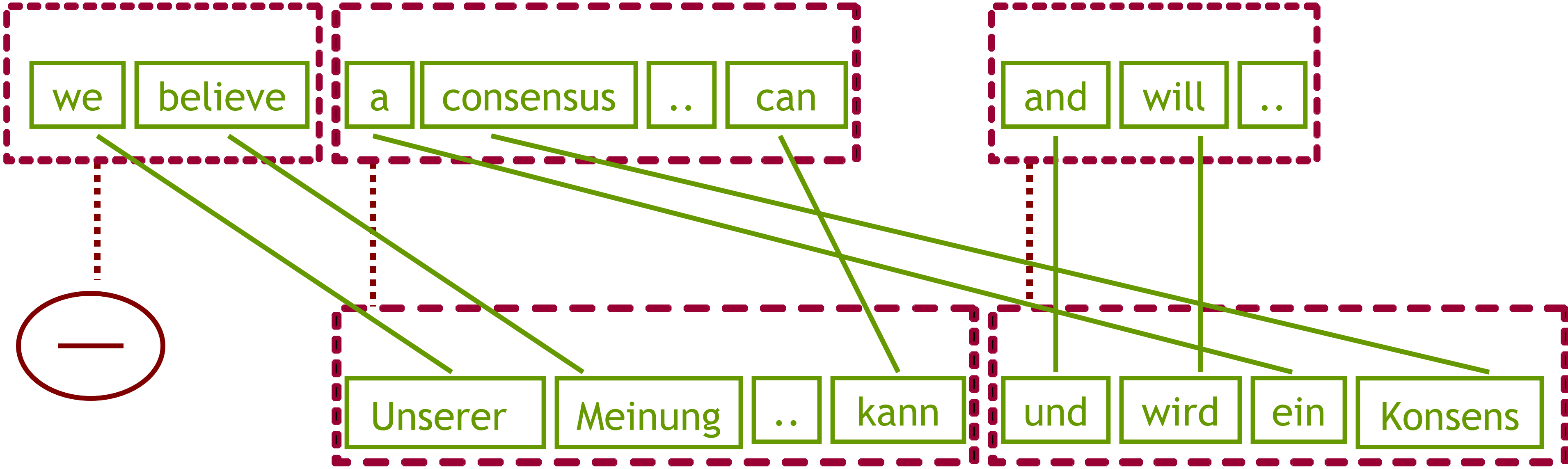


```
<document xmlns:xlink=
 http://www.w3.org/1999/xlink
 name=„GO2Etrans.tokenAlign.xml">
 <translations xml:base="/corpus/">
   <translation trans.loc=„GO.tok.xml„
                     xml:lang="ge" n="1"/>
  <translation trans.loc=„Etrans.tok.xml"
                     xml:lang="en" n="2"/>
 </translations>
 <tokens>
  <token>
   <align xlink:href="#t3"/>
   <align xlink:href="#undefined"/>
  </token>
  <token>
   <align xlink:href="#undefined"/>
   <align xlink:href="#t4"/>
  </token>
  <token>
    <align xlink:href="#t4"/>
   <align xlink:href="#t1"/>
  </token>
</document>
```

XML Clause Alignment

German source

[We believe] [a consensus about Britain's role in Europe can] [and will be built.]



English target

[Unserer Meinung nach kann] [und wird ein Konsens über Großbritanniens Rolle in Europa herbeigeführt werden.]

Query the Corpus for Crossing Lines

```
for $i in doc("eaclq.go2etrans.tokenAlign.xml")//tokens/token
let $tok1:=
    (if ($i/align[1][@xlink:href != "#undefined"] and $i/align[2][@xlink:href != "#undefined"])
            then
            (doc(doc("eaclq.go2etrans.tokenAlign.xml")//translations/translation[@n='1']/
            @trans.loc)//tokens/token[@id eq substring-after($i/align[1]/@xlink:href, "#")])
            else ())
let $tok2:=
     (if ($i/align[1][@xlink:href != "#undefined"] and $i/align[2][@xlink:href != "#undefined"])
            then
            (doc(doc("eaclq.go2etrans.tokenAlign.xml")//translations/translation[@n='2']/
            @trans.loc)//tokens/token[@id eq substring-after($i/align[2]/@xlink:href, "#")])
            else ())
where
            (local:containsToken($ch1/tok[position()=1], $ch1/tok[last()], $tok1/@id) and
not(local:containsToken($ch2/tok[position()=1], $ch2/tok[last()], $tok/@id)))
return $tok1
```

Querying explicitation

XQuery:
Return all units with a PRELS part-of-speech tag which are not aligned on the token level (empty link)

```
for $k in $doc//tokens/token
  let $fileName := $doc//translations/translation[@n='1']/@trans.loc
  let $fileNameNew := replace($fileName,"tok","tag" )
    where ($k/align[1][@xlink:href != "#undefined"] and $k/align[2]
     [@xlink:href = "#undefined"] and doc($fileNameNew)//token
     [@xlink:href eq $k/align[1]/@xlink:href][@pos eq "prels"])
```

Explicitation of pronominal relation
+ participant role
+ tense
+ mood

English original

a palmist, inferring the future out of his own lined flesh

German translation

ein Handleser, der seine Zukunft aus den eigenen Linien ableitete

Outlook

- Corpus access via Internet
- Graphical query interface
- Empirical (+ statistical) analysis of explicitation (and other translation properties)
- Definition of “the translation unit”?