EUROPEAN LANGUAGE GRID

universität wien

Extracting Terminological Concept Systems from Natural Language Text (Text2TCS)

Text2TCS

Terminological Concept Systems (TCS)

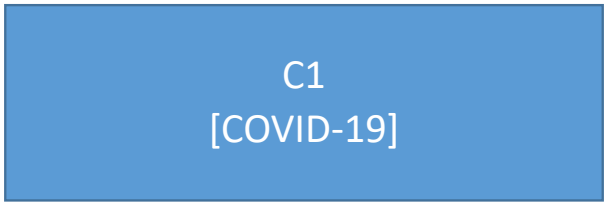
- Concept: abstractions of a set of physical or abstract entities
- Term: their designation by linguistic means
- Relation: hierarchical and semantic relations between concepts that structure and organize the TCS

TCS From Text

Text Coronavirus disease 2019 (COVID-19) is a disease that has been spreading through Europe.

Terms <term>*Coronavirus disease 2019*</term>
<term>*COVID-19*</term> <term>*disease*</term>
<term>*spreading*</term> <term>*Europe*</term>

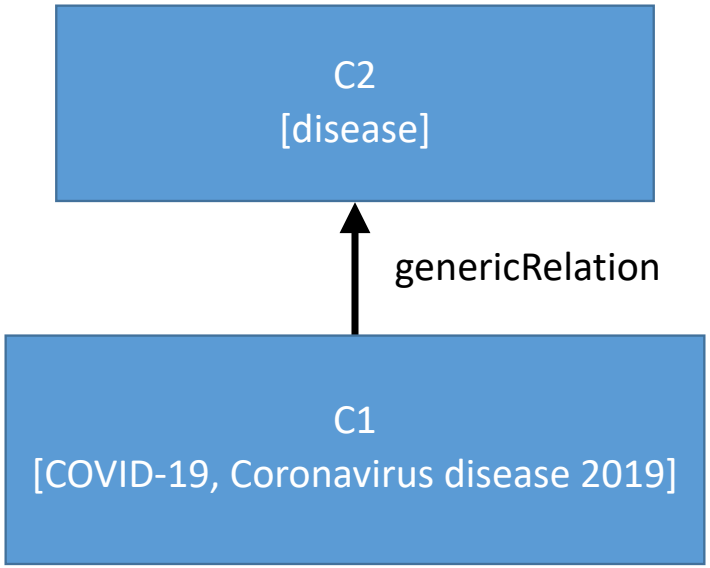
TCS From Text



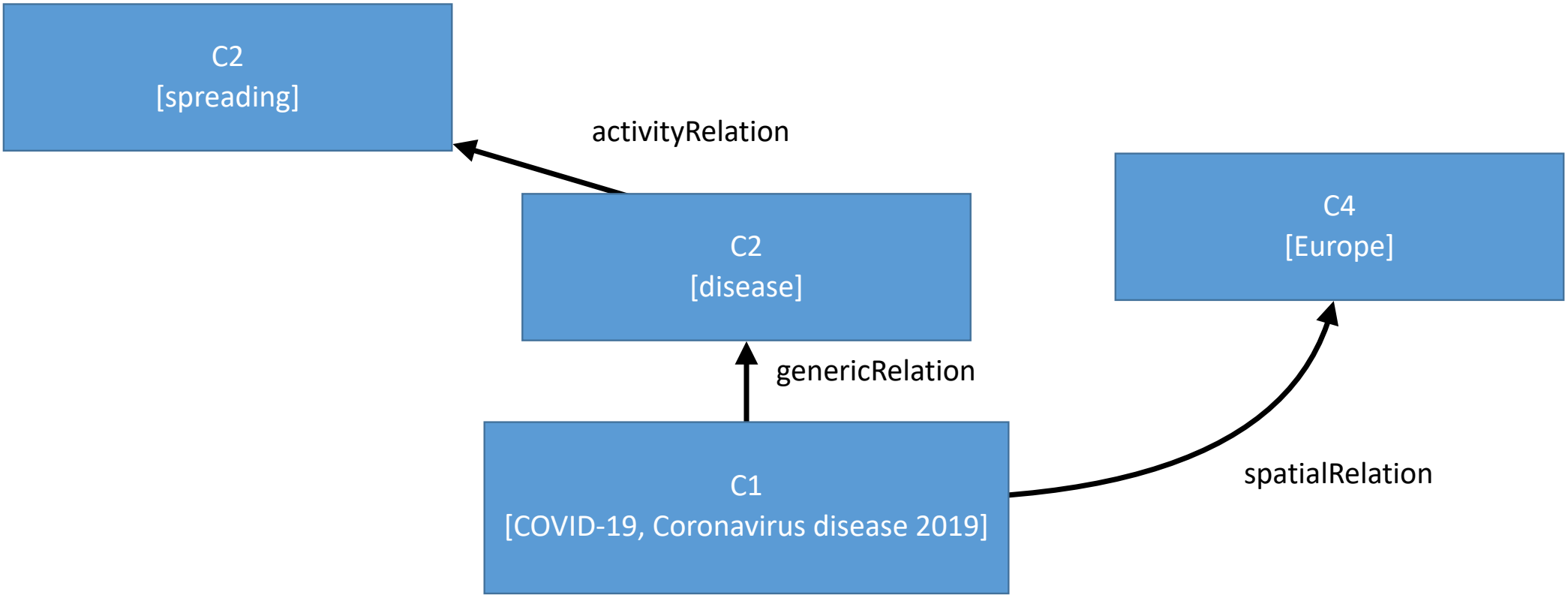
TCS From Text

C1
[COVID-19, Coronavirus disease 2019]

TCS From Text



TCS From Text

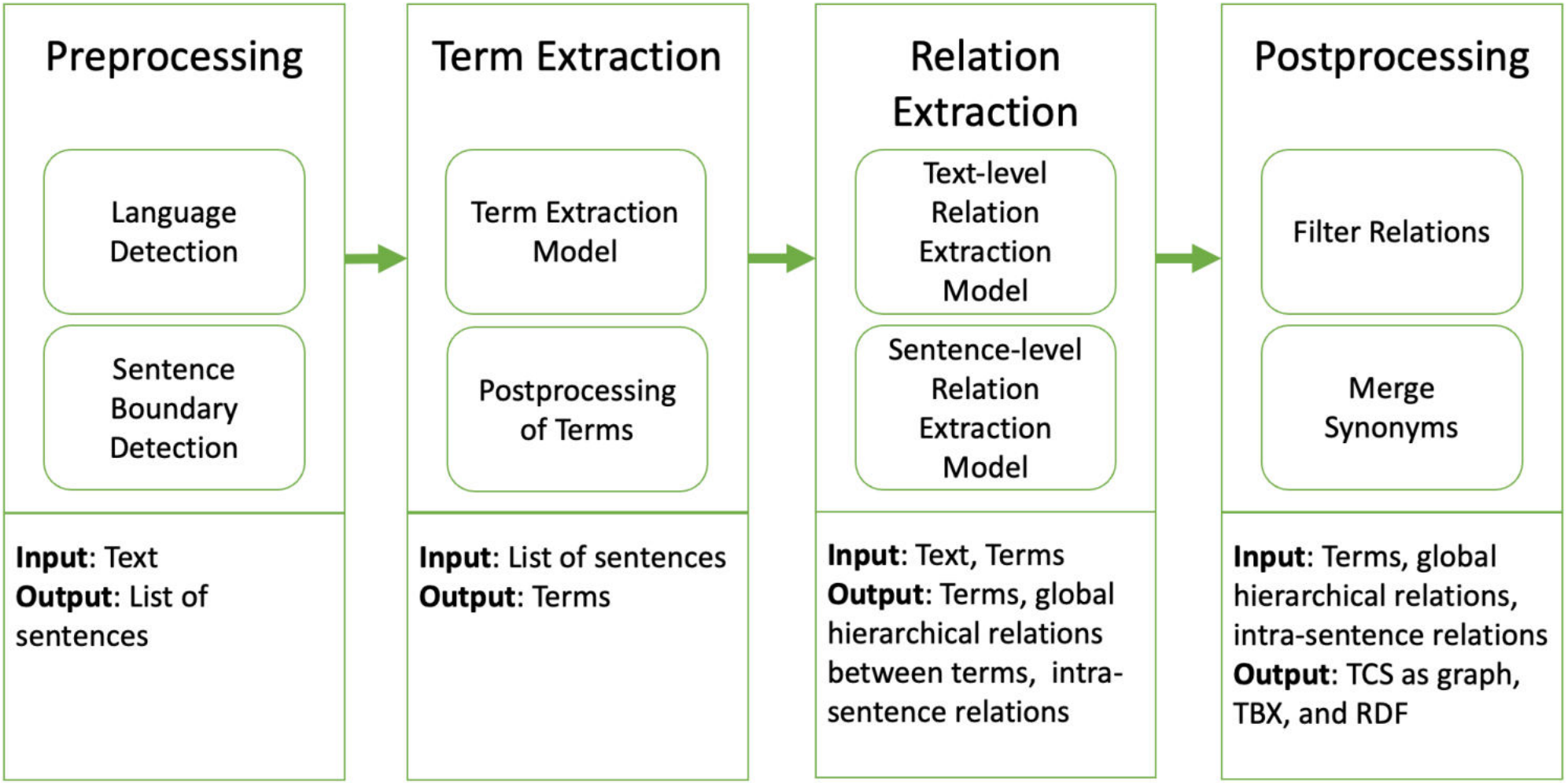


Motivation

- Specialized communication needs to be able to rely on a consistent usage of terminology by different parties
- A TCS alleviates that problem, but manually curating a TCS is cumbersome

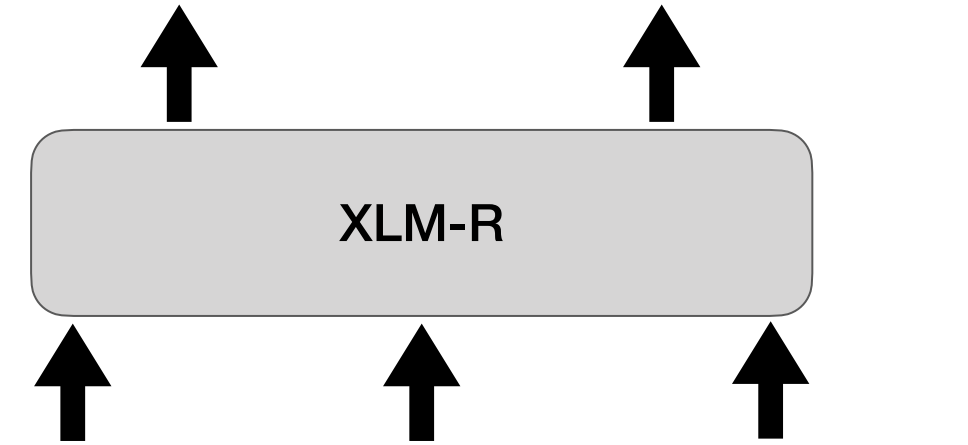
→ **Automated approaches are needed**

Text2TCS — the Pipeline

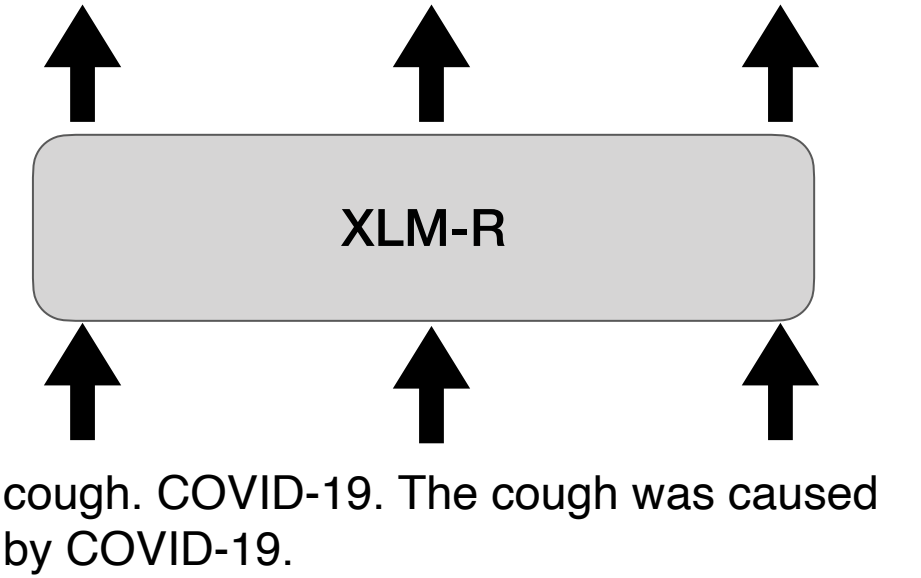


Method - Multilingual BERT models (XLM-R)

Term Extraction



Relation Extraction



Lang C. et al (2021) Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.

Wachowiak L. et al. (2021) Towards Learning Terminological Concept Systems from Multilingual Natural Language Text. in LDK 2021

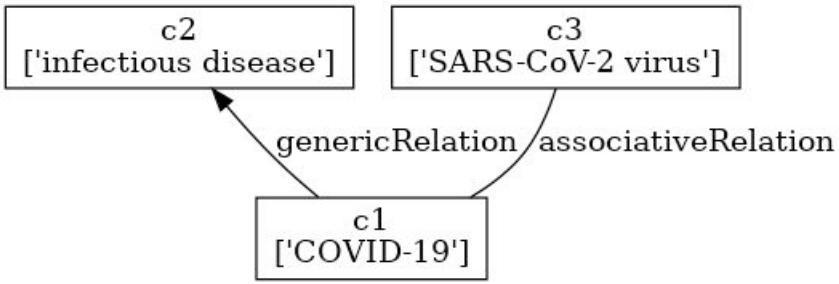
Text2TCS on ELG

https://live.european-language-grid.eu/catalogue/tool-service/8122

Text2TCS on ELG

https://live.european-language-grid.eu/catalogue/tool-service/8122

Concept Map



TBX/XML

```
<?xml version='1.0' encoding='utf8'?>
<?xml-model href="https://raw.githubusercontent.com/LTAC-Global/TBX-Core_dialect/master/Schemas/TBX
schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model href="https://raw.githubusercontent.com/LTAC-Global/TBX-Core_dialect/master/Schemas/TBX
/dsdl/schematron"?>
<tbx type="TBX-Core" style="dca" xml:lang="en" xmlns="urn:iso:std:iso:30042:ed-2">
  <tbxHeader>
    <fileDesc>
      <sourceDesc>
        <p>TBX file automatically generated by Text2TCS (https://text2tcs.univie.ac.at/)</p>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <p type="XCSURI">TBXXCSV02.xcs</p>
    </encodingDesc>
  </tbxHeader>
  <text>
    <body>
      <conceptEntry id="c1">
        <transacGrp>
          <transac type="transactionType">origination</transac>
          <transacNote type="responsibility">Text2TCS</transacNote>
          <date>21-11-11_17h-34m</date>
        </transacGrp>
        <langSec xml:lang="en">
          <termSec id="c1-en-t0">
            <term>COVID-19</term>
          </termSec>
        </langSec>
        <descripGrp>
          <descrip type="genericRelation">c2</descrip>
        </descripGrp>
        <descripGrp>
          <descrip type="associativeRelation">c3</descrip>
        </descripGrp>
      </conceptEntry>
      <conceptEntry id="c2">
        <transacGrp>
          <transac type="transactionType">origination</transac>
          <transacNote type="responsibility">Text2TCS</transacNote>
          <date>21-11-11_17h-34m</date>
        </transacGrp>
        <langSec xml:lang="en">
          <termSec id="c2-en-t0">
            <term>infectious disease</term>
          </termSec>
        </langSec>
      </conceptEntry>
      <conceptEntry id="c3">
        <transacGrp>
```

Major findings

- Fine-tuning pre-trained neural language models works well on this TCS extraction task
- Works also well with Neural Machine Translation models - even for joint term and relation extraction (but too slow and big for ELG)
- Coordination between sentence- and text-level output challenging - which model to trust more?
- Pre-annotated data still needed
- All depends on good team work…

TEAM

DAGMAR GROMANN
Project leader

Text2TCS

LENNART WACHOWIAK
Machine learning and IT

CHRISTIAN LANG
Translation and IT

BARBARA HEINISCH
Terminology and usability

universität wien

EUROPEAN LANGUAGE GRID

Text2TCS

Thank you for listening!

More information on:

https://text2tcs.univie.ac.at/

Text2TCS on ELG:

https://live.european-language-grid.eu/catalogue/tool-service/8122

Any feedback very welcome:

dagmar.gromann@univie.ac.at