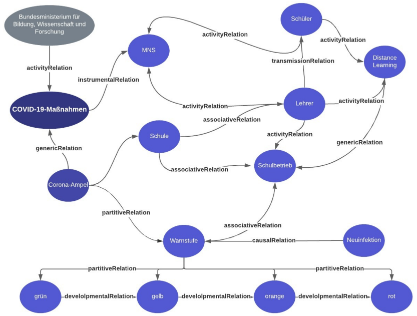
universität
wien

Text2TCS

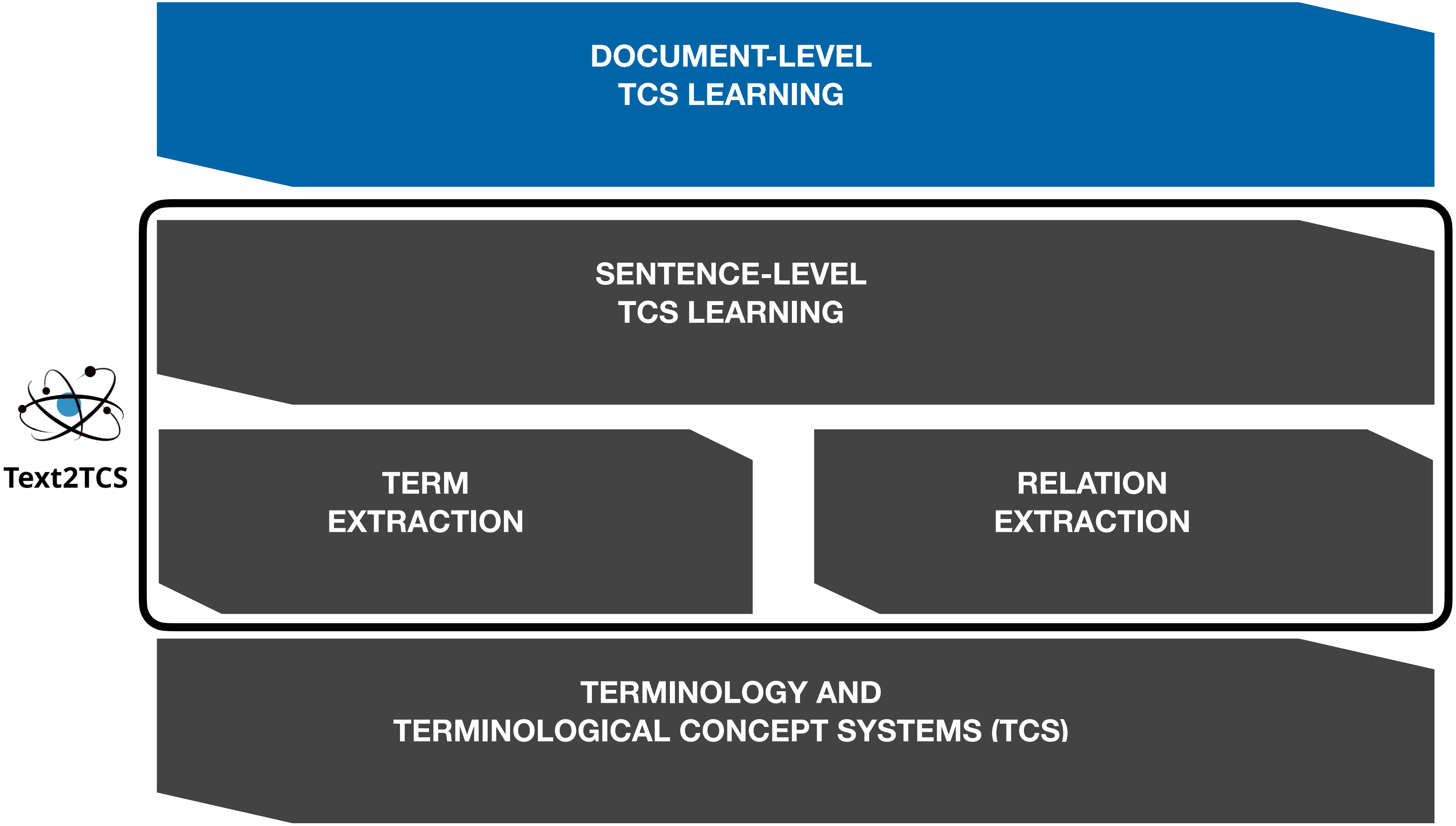
Multilingual Information Extraction to Learn Terminological Concept Systems

Dagmar Gromann, Centre for Translation Studies, University of Vienna
dagmar.gromann@univie.ac.at

Cardiff NLP Seminar, 18 March 2021



Overview

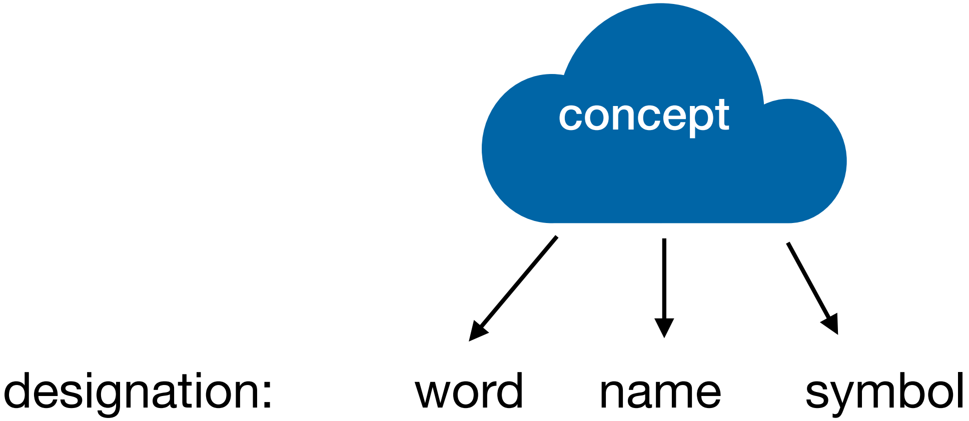


DOCUMENT-LEVEL
TCS LEARNING
SENTENCE-LEVEL
TCS LEARNING
Text2TCS
TERM
EXTRACTION
RELATION
EXTRACTION
TERMINOLOGY AND
TERMINOLOGICAL CONCEPT SYSTEMS (TCS)

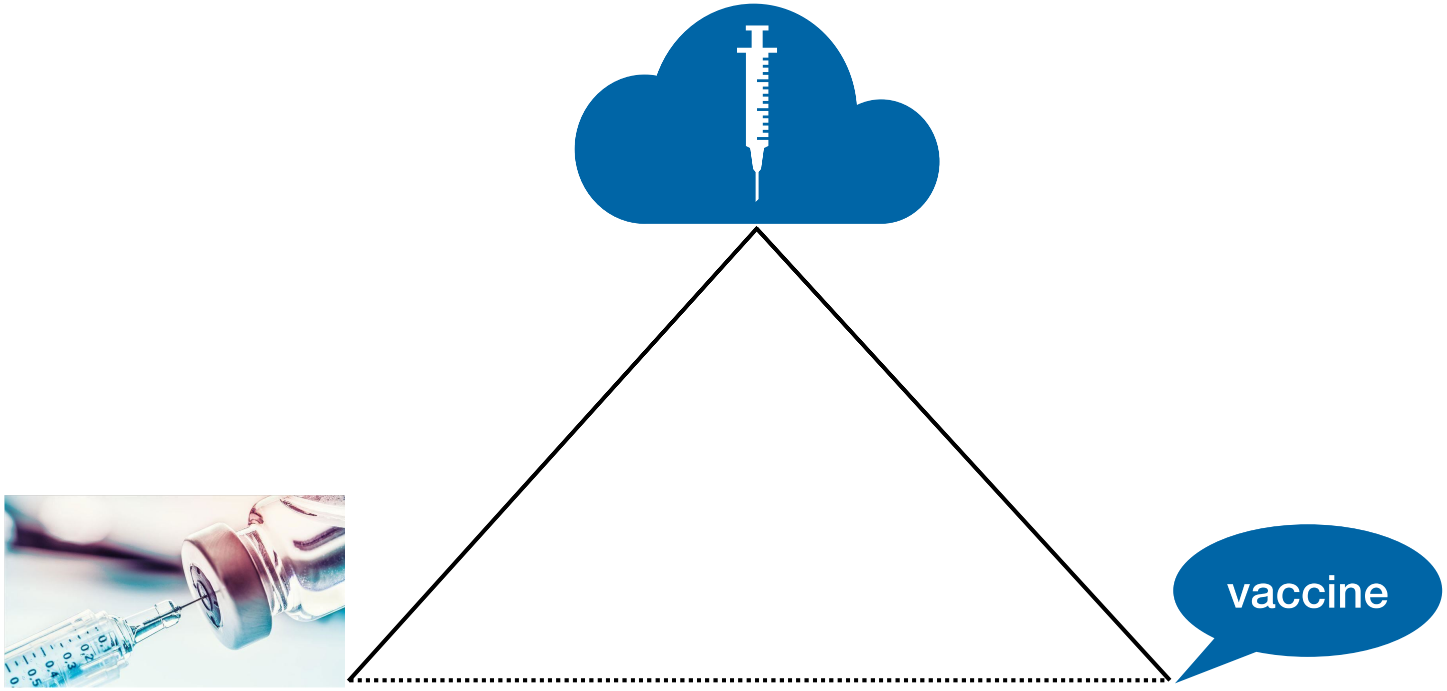
Terminology

- *as a field*: multidisciplinary field of study that borrows from logic, epistemology, linguistics, philosophy, translation studies, and cognitive science

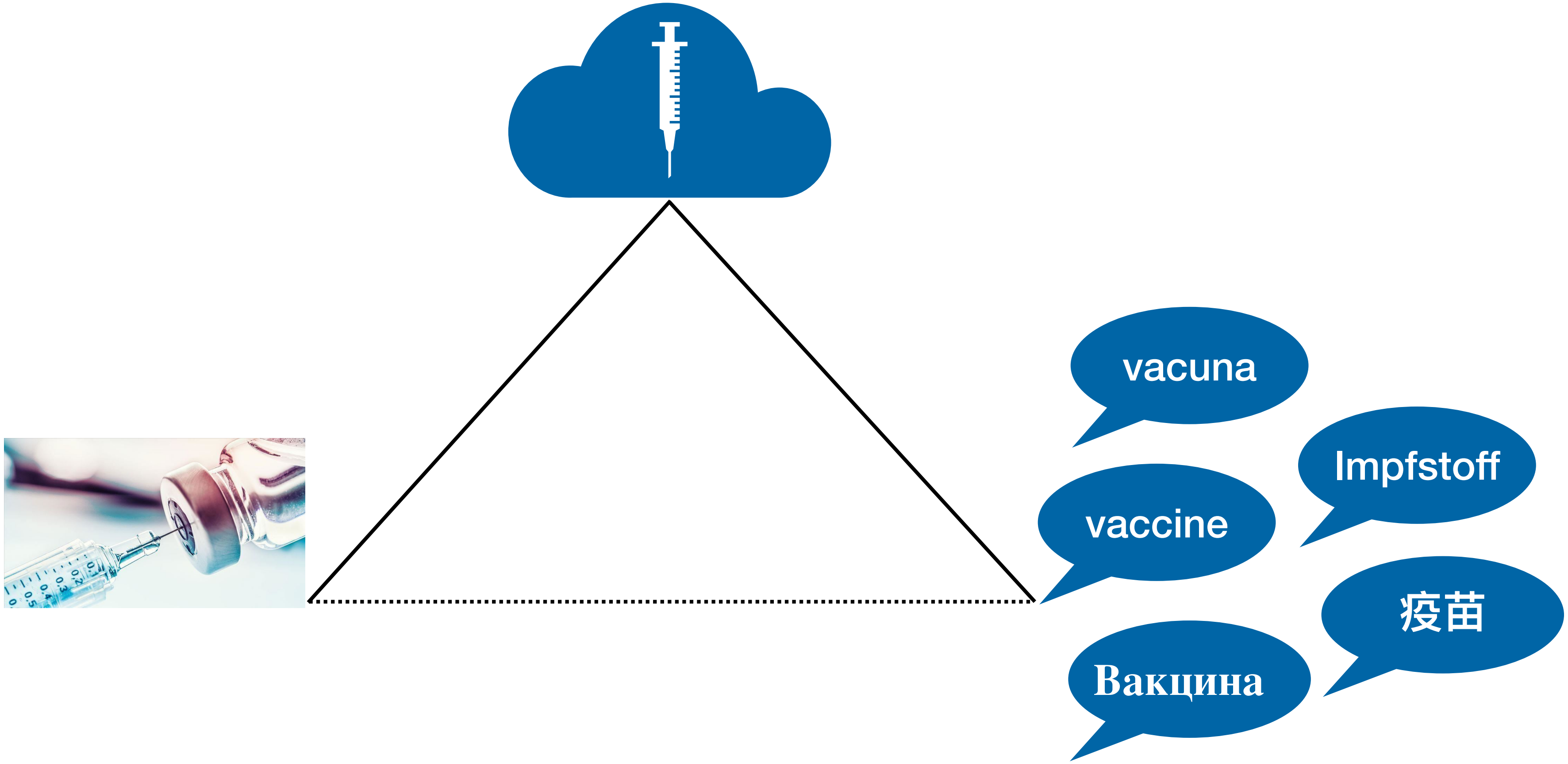
- *as a resource:*
  - collection of concepts, their interrelations and designations in a specialized field
  - multilingual designations in a specific subject field structured by concepts, i.e. domain-specific
  - concept? unit of thought, unit of understanding, unit of specialized communication, …



Terminology



Terminology



vacuna

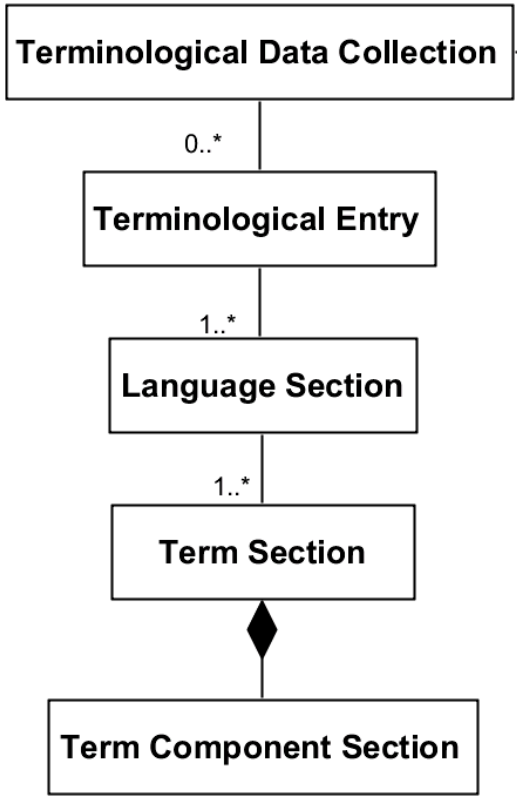
Impfstoff

vaccine

疫苗

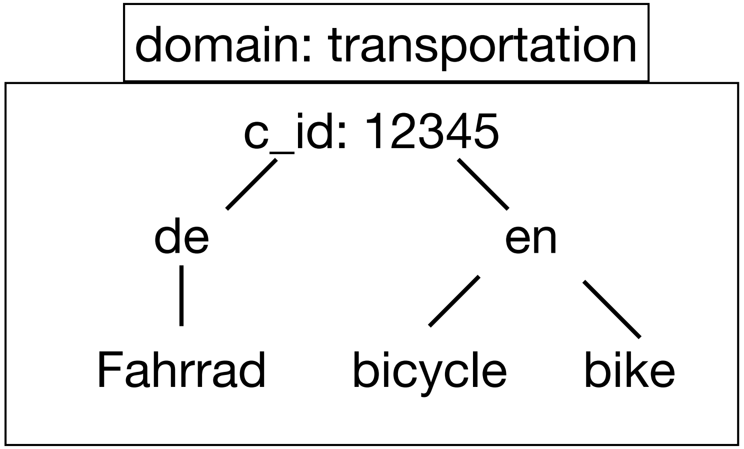
Вакцина

Terminological Entry



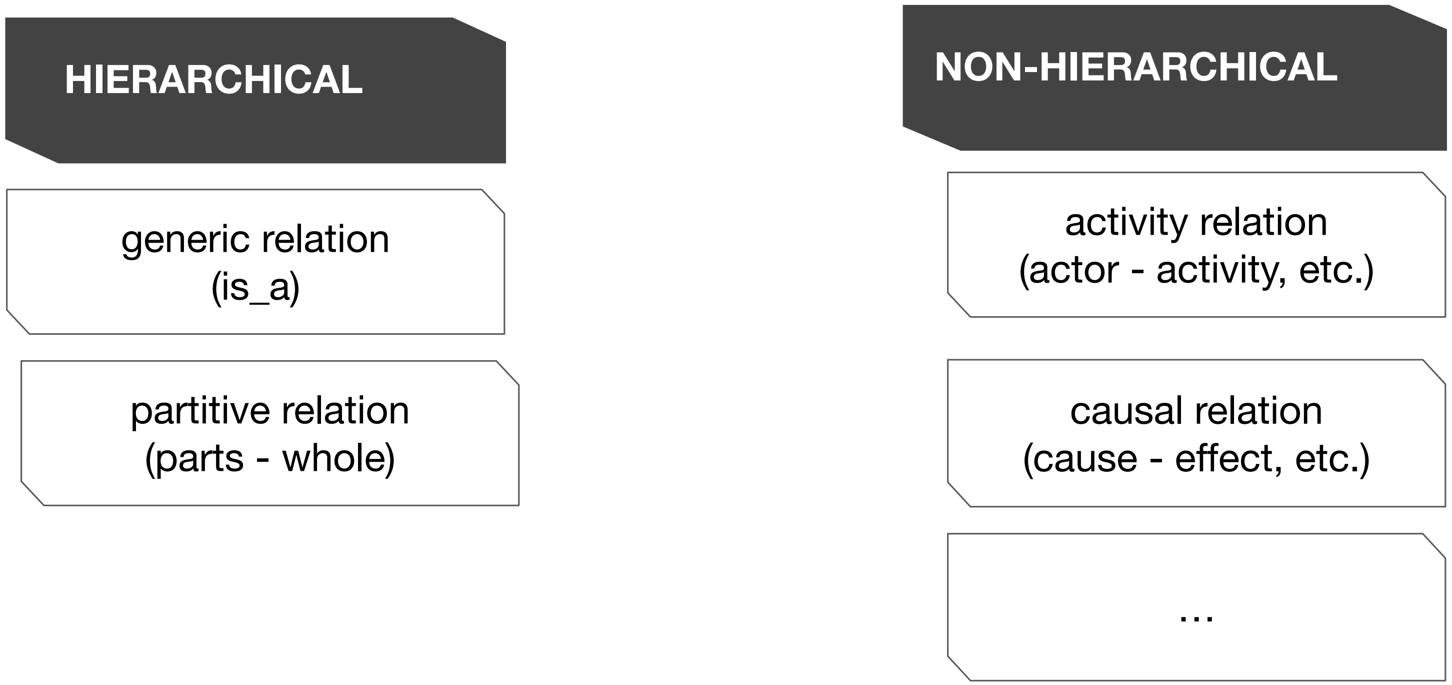
- Basic model of the Terminological Markup Framework (TMF; ISO 16642: 2017)

- nested internal organization of language information in relation to a concept



Terminological Concept Systems (TCS)

- grouping synonyms and equivalents by concept
- interrelating concepts with hierarchical and non-hierarchical relations



HIERARCHICAL

generic relation (is_a)

partitive relation (parts - whole)

NON-HIERARCHICAL

activity relation (actor - activity, etc.)

causal relation (cause - effect, etc.)

…

Primary functions of a TCS

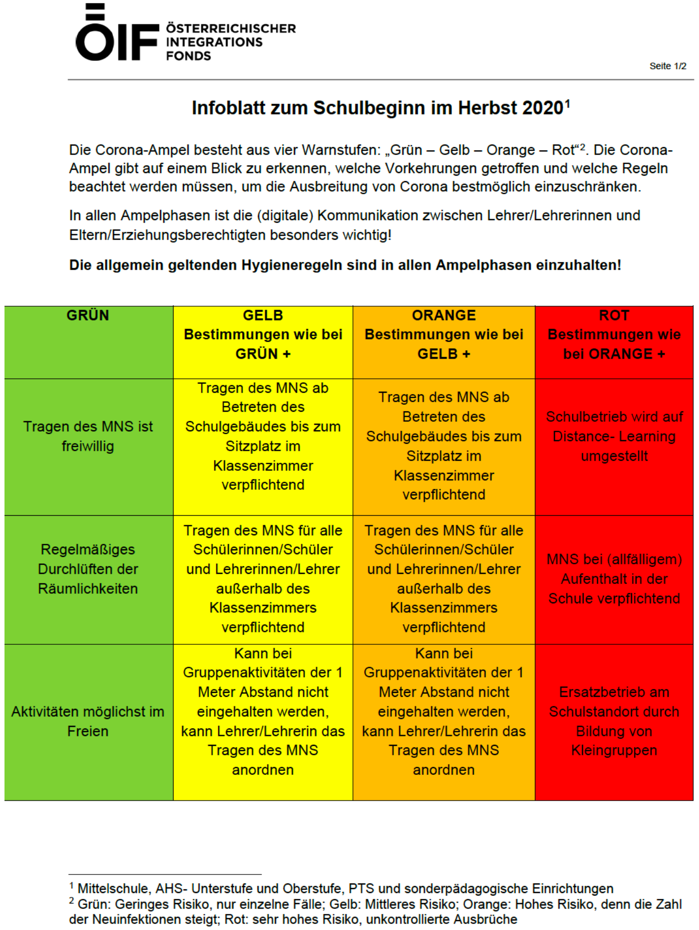
TCS in principle can be described as structuring means with three major functions (Budin 1996: 18):

1. *epistemic*: epistemological instrument in the sense of structuring knowledge (acquire new knowledge)

2. *informational*: structuring means for practical knowledge transfer (structuring existing knowledge)

3. *communicative:* optimization of specialized communication in the sense of communication organization and consistency (extend existing knowledge)

From Text to TCS



ÖSTERREICHISCHER INTEGRATIONS FONDS

Seite 1/2

Infoblatt zum Schulbeginn im Herbst 2020[1]

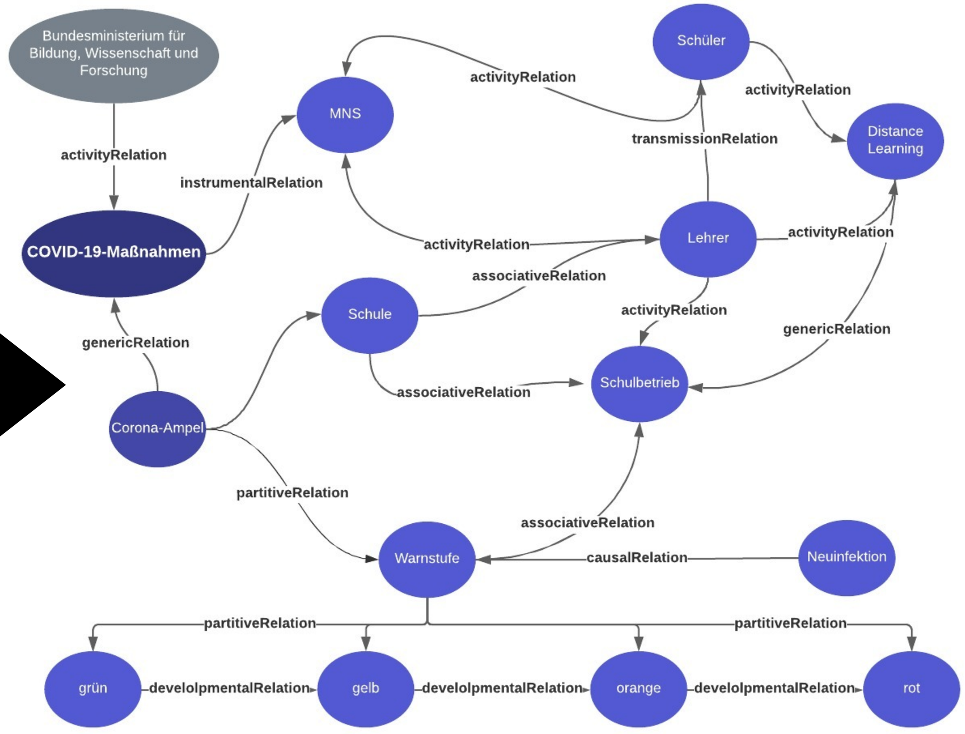
Die Corona-Ampel besteht aus vier Warnstufen: „Grün – Gelb – Orange – Rot“[2]. Die Corona-Ampel gibt auf einem Blick zu erkennen, welche Vorkehrungen getroffen und welche Regeln beachtet werden müssen, um die Ausbreitung von Corona bestmöglich einzuschränken.

In allen Ampelphasen ist die (digitale) Kommunikation zwischen Lehrer/Lehrerinnen und Eltern/Erziehungsberechtigten besonders wichtig!

Die allgemein geltenden Hygieneregeln sind in allen Ampelphasen einzuhalten!

GRÜN	GELB Bestimmungen wie bei GRÜN +	ORANGE Bestimmungen wie bei GELB +	ROT Bestimmungen wie bei ORANGE +
Tragen des MNS ist freiwillig	Tragen des MNS ab Betreten des Schulgebäudes bis zum Sitzplatz im Klassenzimmer verpflichtend	Tragen des MNS ab Betreten des Schulgebäudes bis zum Sitzplatz im Klassenzimmer verpflichtend	Schulbetrieb wird auf Distance- Learning umgestellt
Regelmäßiges Durchlüften der Räumlichkeiten	Tragen des MNS für alle Schülerinnen/Schüler und Lehrerinnen/Lehrer außerhalb des Klassenzimmers verpflichtend	Tragen des MNS für alle Schülerinnen/Schüler und Lehrerinnen/Lehrer außerhalb des Klassenzimmers verpflichtend	MNS bei (allfälligem) Aufenthalt in der Schule verpflichtend
Aktivitäten möglichst im Freien	Kann bei Gruppenaktivitäten der 1 Meter Abstand nicht eingehalten werden, kann Lehrer/Lehrerin das Tragen des MNS anordnen	Kann bei Gruppenaktivitäten der 1 Meter Abstand nicht eingehalten werden, kann Lehrer/Lehrerin das Tragen des MNS anordnen	Ersatzbetrieb am Schulstandort durch Bildung von Kleingruppen

[1] Mittelschule, AHS- Unterstufe und Oberstufe, PTS und sonderpädagogische Einrichtungen
[2] Grün: Geringes Risiko, nur einzelne Fälle; Gelb: Mittleres Risiko; Orange: Hohes Risiko, denn die Zahl der Neuinfektionen steigt; Rot: sehr hohes Risiko, unkontrollierte Ausbrüche



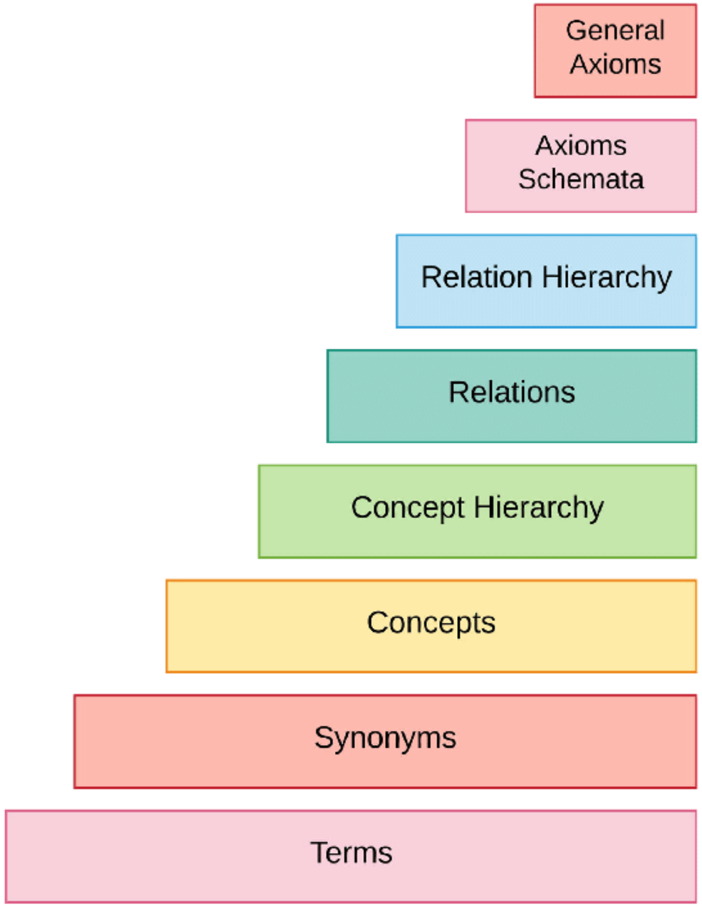
Related Fields

- Ontology learning: e.g. Petrucci et al. (2018) utilize Neural Machine Translation (NMT)

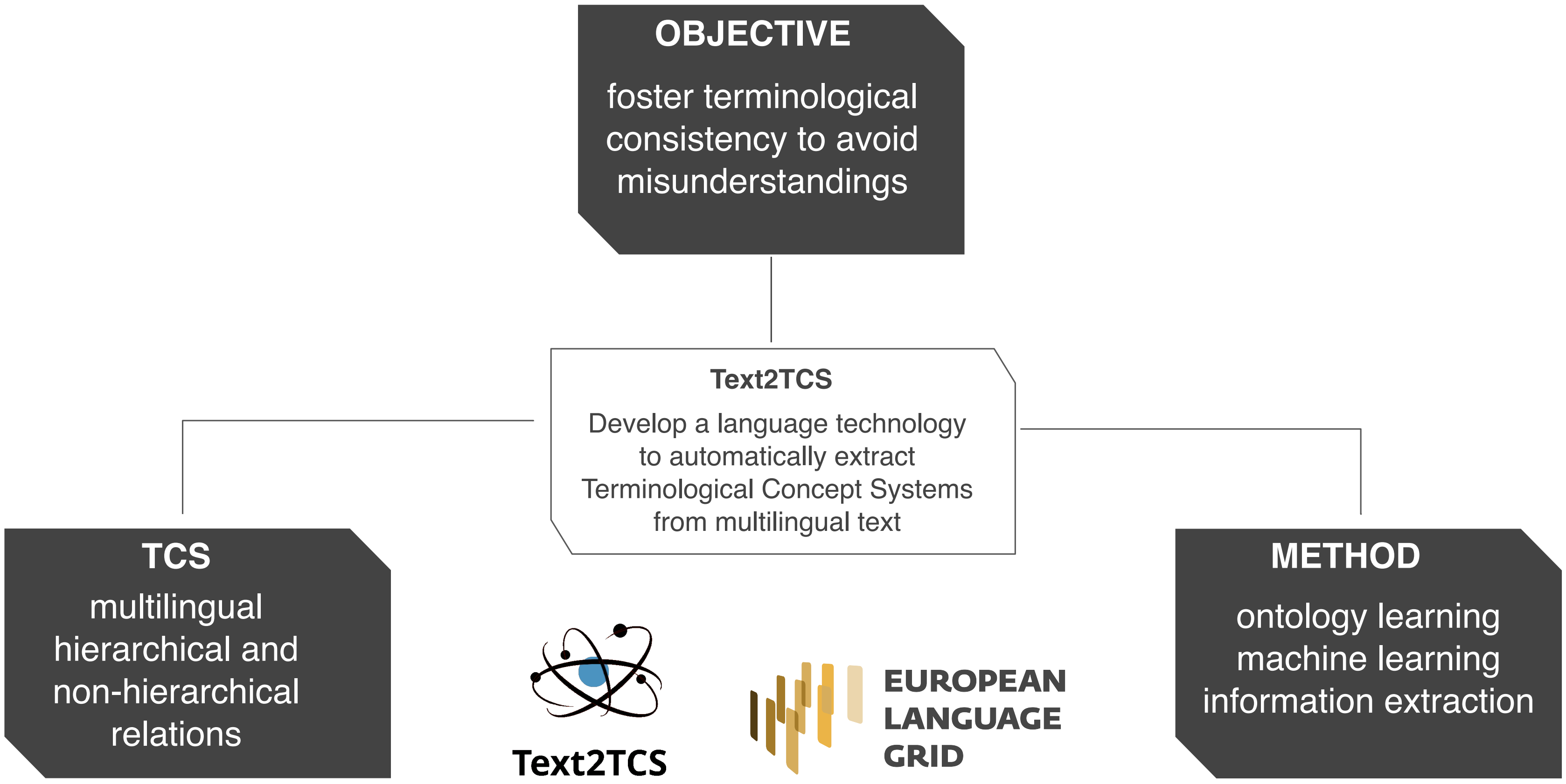
- Entity extraction and linking:

    - interlinking named entities with semantic, (non-)hierarchical relations

    - mostly on sentence-level, some exceptions on document-level (e.g. DocRED by Yao et al. 2019)



Extracting Terminological Concept Systems from Natural Language Text (Text2TCS)



OBJECTIVE

foster terminological consistency to avoid misunderstandings

Text2TCS

Develop a language technology to automatically extract Terminological Concept Systems from multilingual text

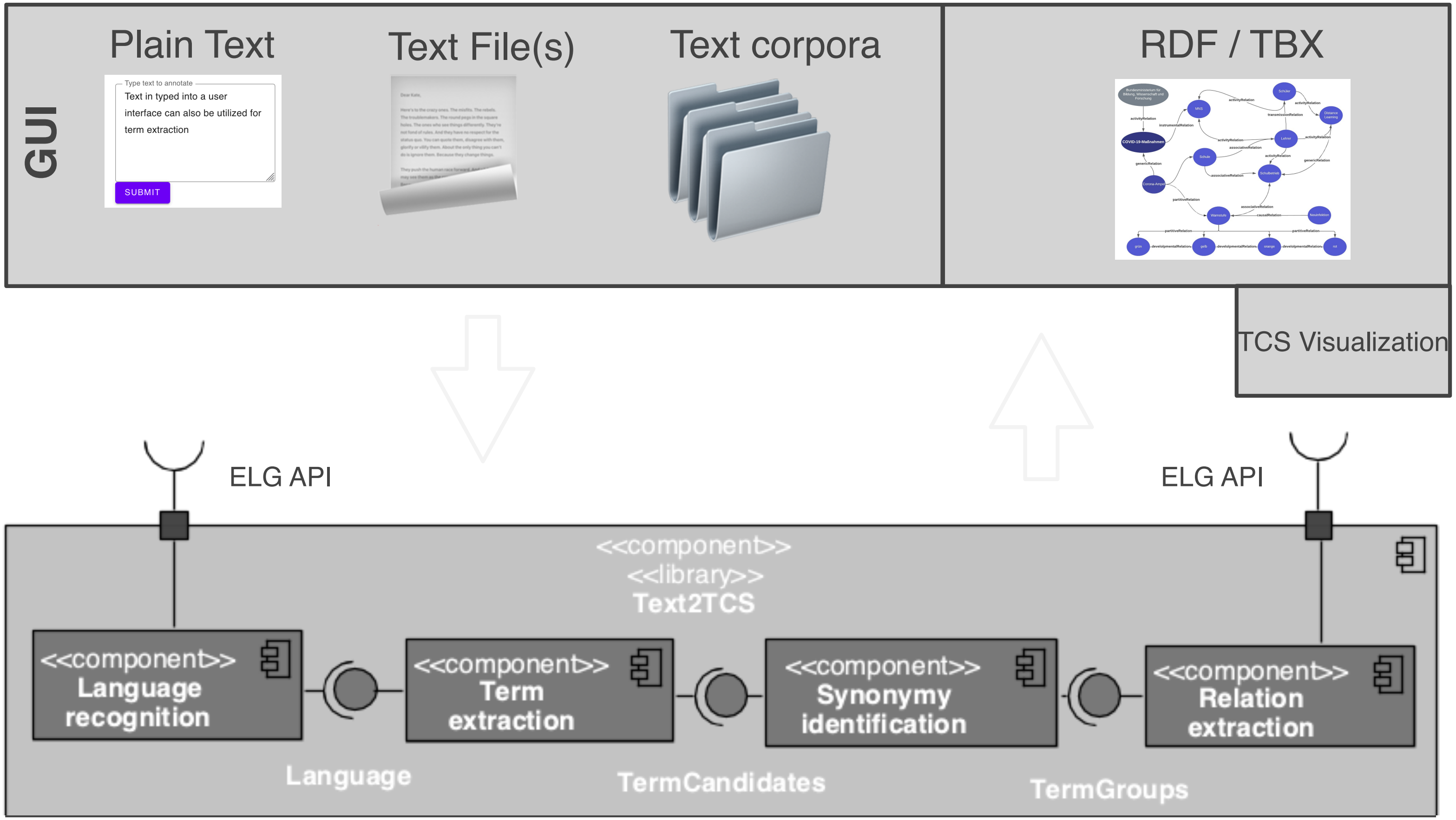
TCS

multilingual hierarchical and non-hierarchical relations

METHOD

ontology learning machine learning information extraction

Text2TCS Components



Type text to annotate

Text in typed into a user interface can also be utilized for term extraction

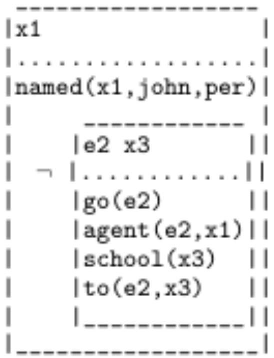
Initial Idea

Dear Kate,

Here's to the crazy ones. The misfits. The rebels. The troublemakers. The round pegs in the square holes. The ones who see things differently. They're not fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things.

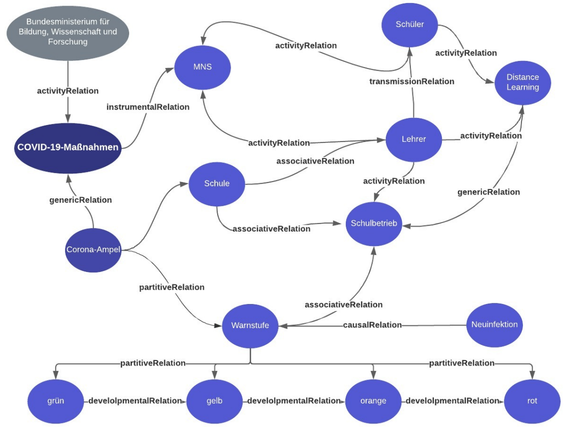
They push the human race forward. And
may see them as the

DRS



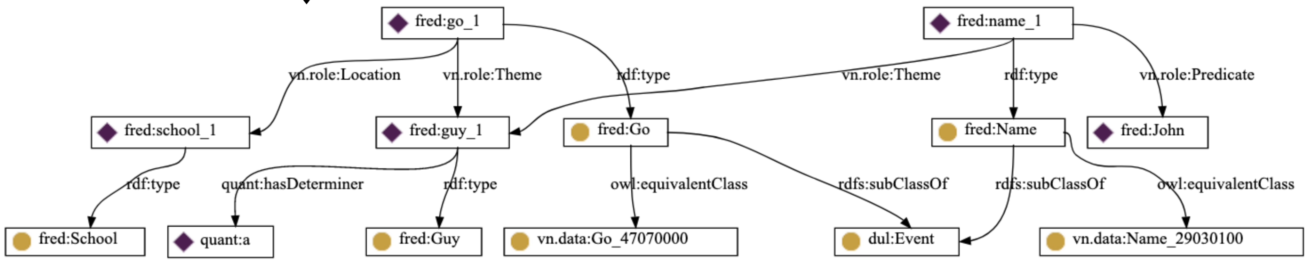
```
 -------------------------
|x1                       |
|.........................|
|named(x1,john,per)       |
|                         |
|     ---------------     |
|    |e2 x3          |    |
|  ¬ |...............|    |
|    |go(e2)         |    |
|    |agent(e2,x1)   |    |
|    |school(x3)     |    |
|    |to(e2,x3)      |    |
|    |_______________|    |
|_________________________|
```

TCS



FRED

Machine Reading for the Semantic Web



http://wit.istc.cnr.it/stlab-tools/fred/

Initial Idea

Dear Kate,

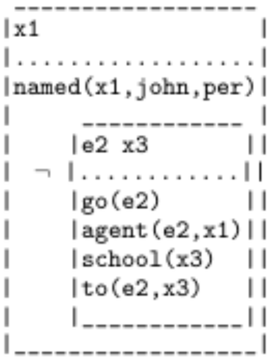
Here's to the crazy ones. The misfits. The rebels. The troublemakers. The round pegs in the square holes. The ones who see things differently. They're not fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things.

They push the human race forward. And

may see them as the

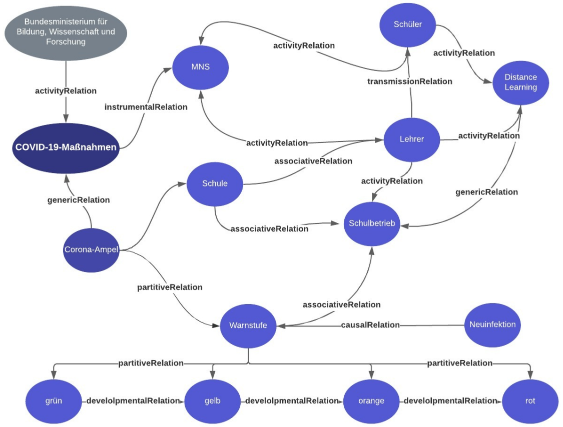
Beca

DRS

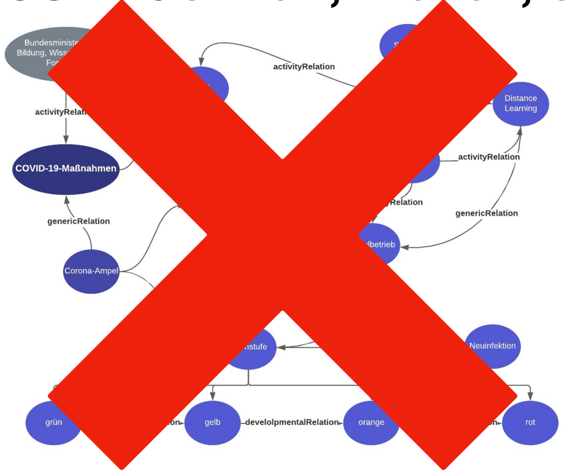


```
 -------------------
|x1                 |
|...................|
|named(x1,john,per) |
|                   |
|    -------------  |
|   |e2 x3        | |
| ¬ |.............| |
|   |go(e2)       | |
|   |agent(e2,x1) | |
|   |school(x3)   | |
|   |to(e2,x3)    | |
|   |_____________| |
|___________________|
```

TCS

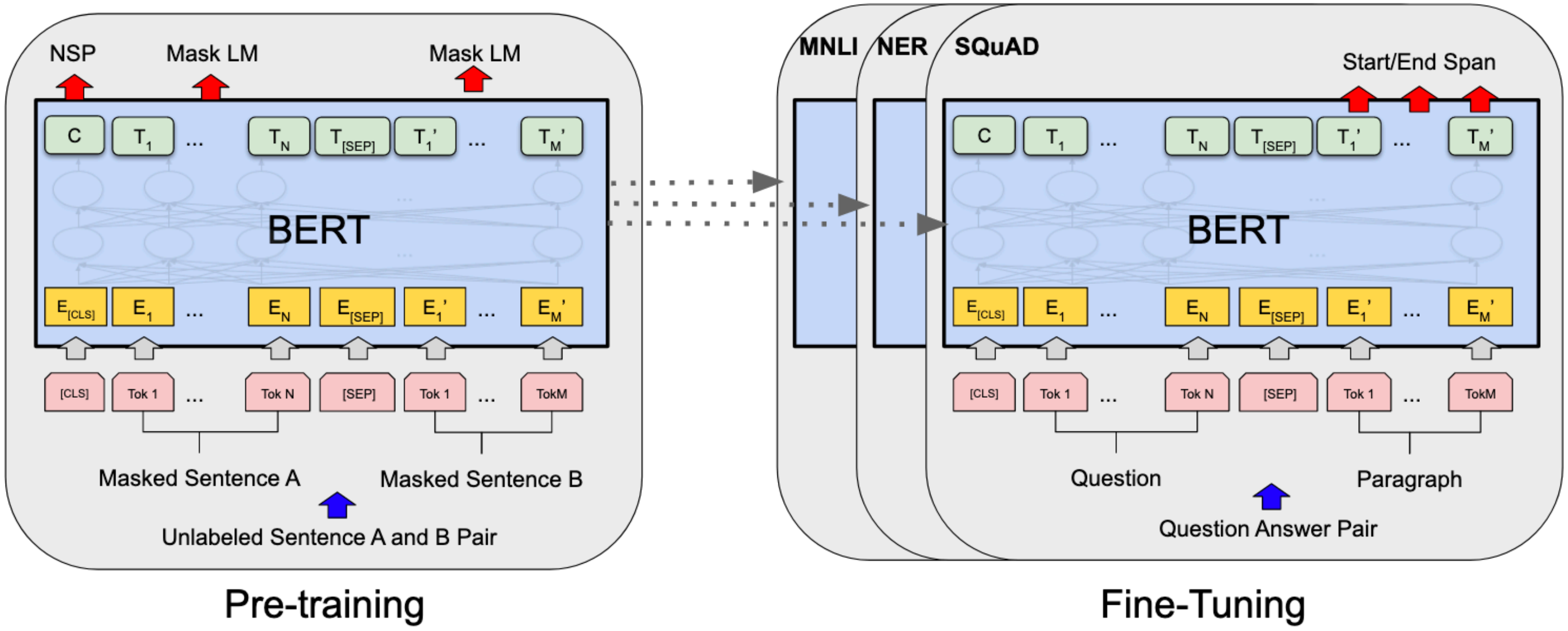


TCS in German, French, etc.



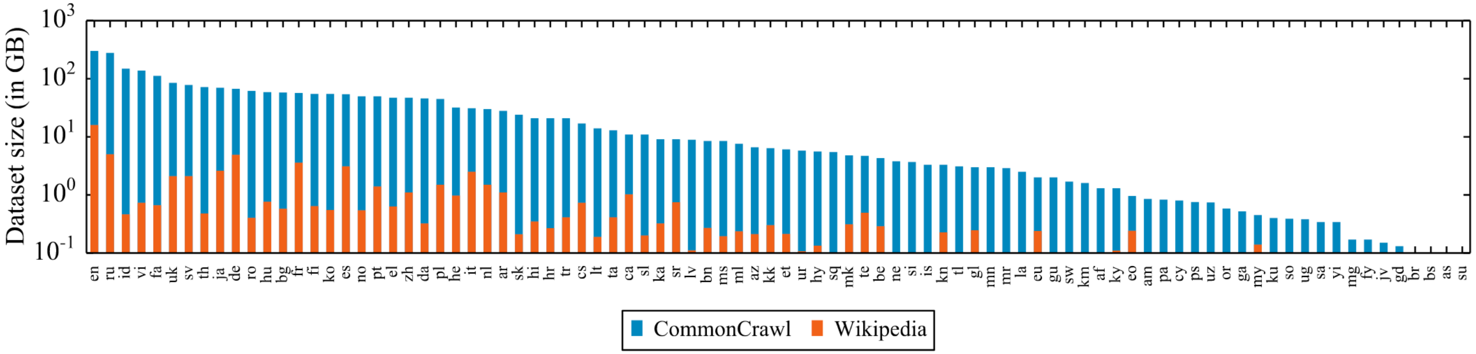
Pretrained Multilingual Language Models

Bidirectional Encoder Representations from Transformers (BERT)



Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

XLM-R



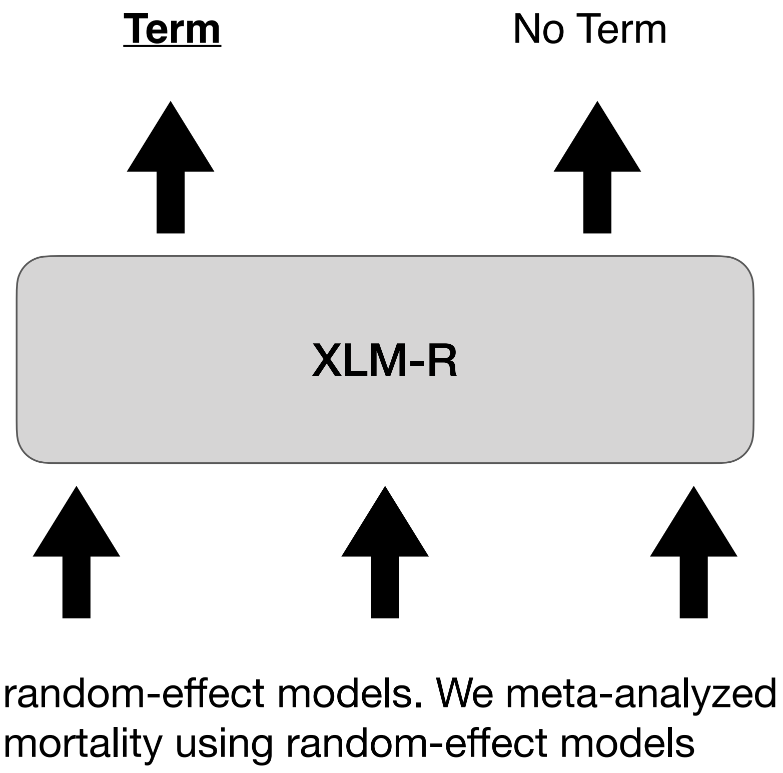
Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. ACL 2020.

LM-based TCS Learning

- Pretrained Language Models (LMs) can easily be adapted to a specific task
- Task at hand: Given a sentence identify all domain-specific terms and relations between them
- Challenges:
    - n-gram length that combines to form a term?
    - domain-specific, e.g. *vaccine* vs. cross-domain-specific, e.g. statistics - both should be extracted
    - Restrictions on sequence length and ability to ensure that two related terms occur in the same input sequence - difficult across sentences but also within long sentences
    - enable this task across many different languages

Term Extraction

Sequence-Classifier

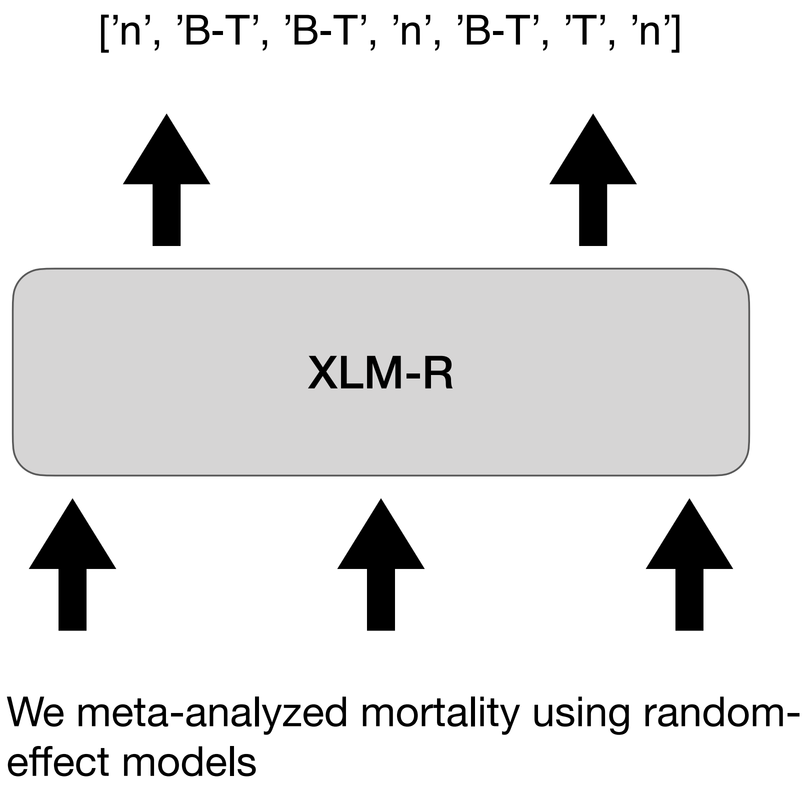


<u>Term</u> No Term

XLM-R

random-effect models. We meta-analyzed mortality using random-effect models

Token-Classifier



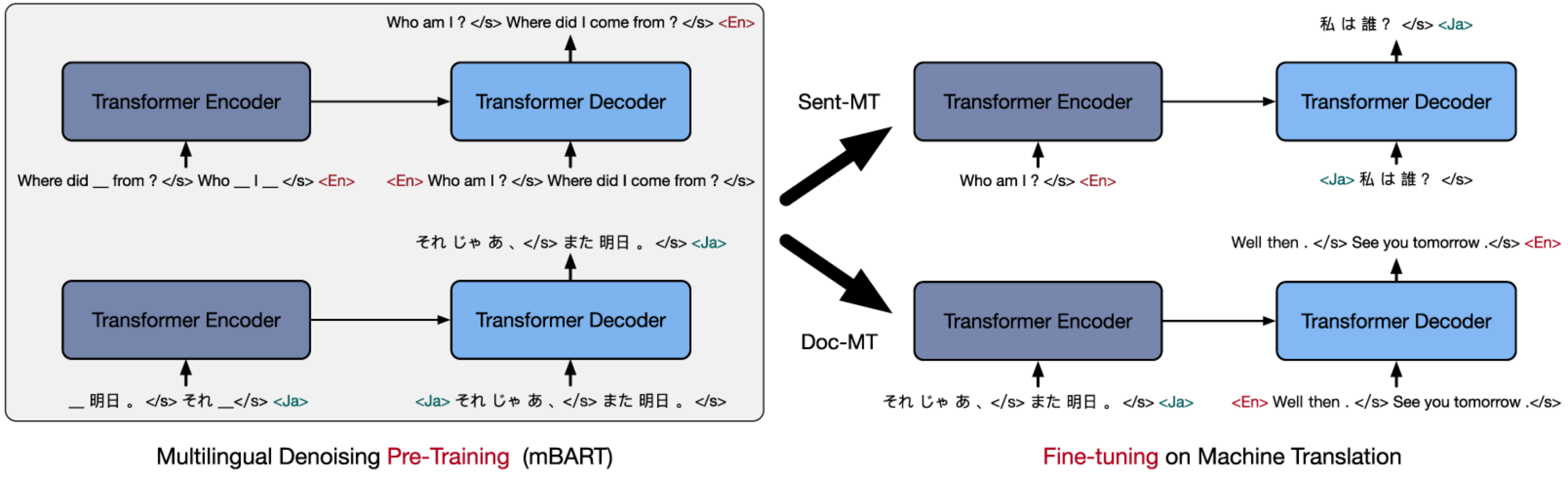
[’n’, ’B-T’, ’B-T’, ’n’, ’B-T’, ’T’, ’n’]

XLM-R

We meta-analyzed mortality using random-effect models

Machine Translating Term Extraction with mBART

meta-analyzed ; mortality ; random-effect models


Who am I ? </s> Where did I come from ? </s> <En>
Transformer Encoder
Transformer Decoder
Where did __ from ? </s> Who __ I __ </s> <En>
<En> Who am I ? </s> Where did I come from ? </s>
それ じゃ あ 、</s> また 明日 。 </s> <Ja>
Transformer Encoder
Transformer Decoder
__ 明日 。 </s> それ __</s> <Ja>
<Ja> それ じゃ あ 、</s> また 明日 。 </s>
Multilingual Denoising Pre-Training (mBART)
Sent-MT
Doc-MT
私 は 誰 ？ </s> <Ja>
Transformer Encoder
Transformer Decoder
Who am I ? </s> <En>
<Ja> 私 は 誰 ？ </s>
Well then . </s> See you tomorrow .</s> <En>
Transformer Encoder
Transformer Decoder
それ じゃ あ 、</s> また 明日 。 </s> <Ja>
<En> Well then . </s> See you tomorrow .</s>
Fine-tuning on Machine Translation

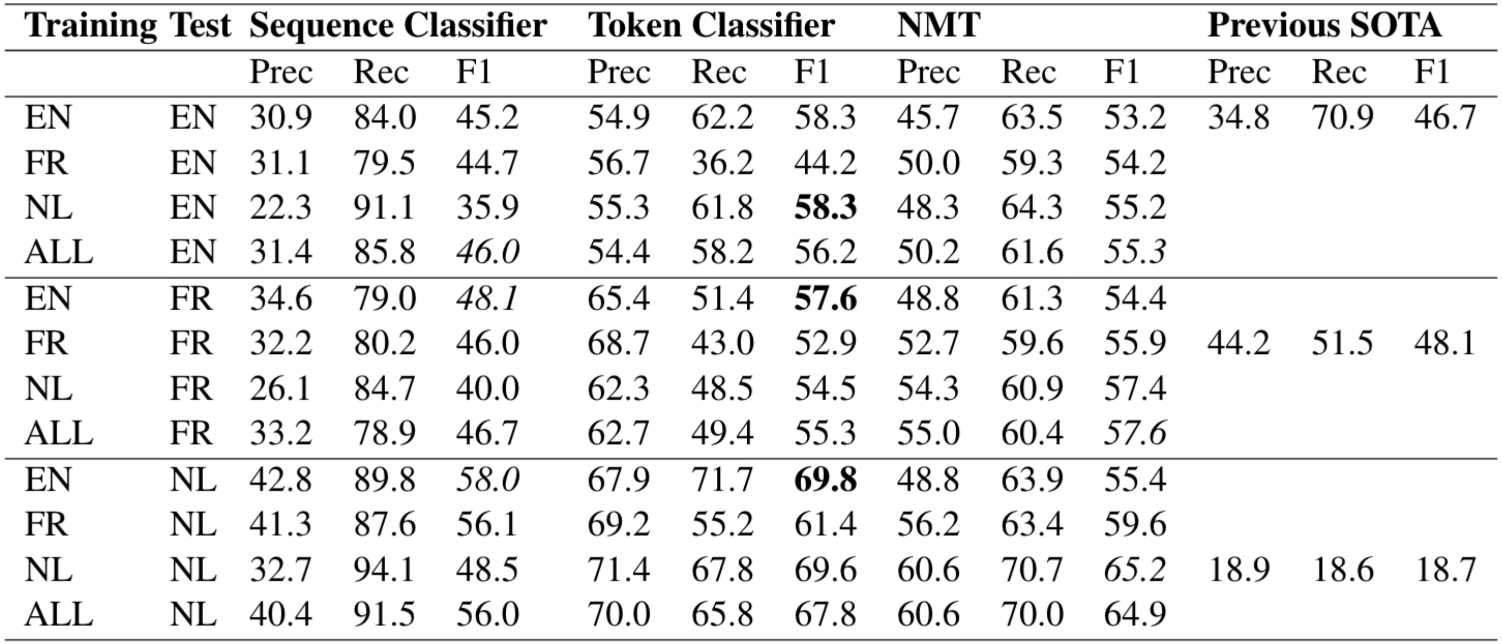
We meta-analyzed mortality using random-effect models.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726-742.

Results on TermEval

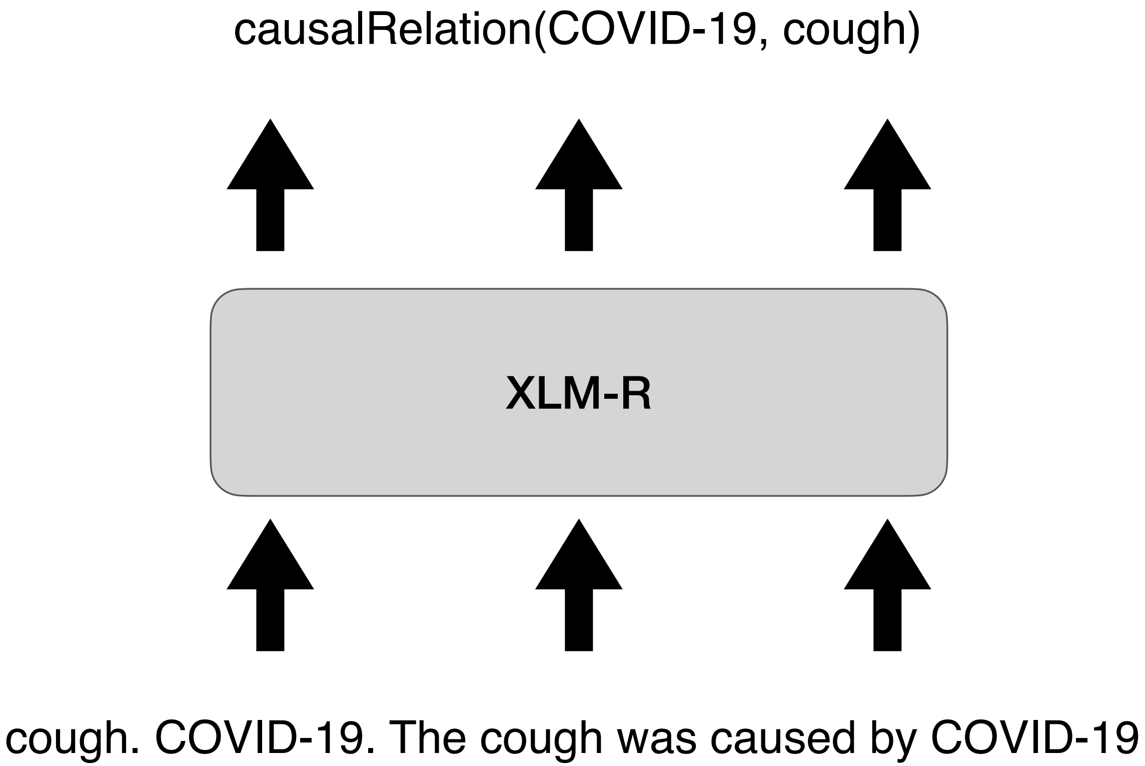
TermEval (Rigouts et al. 2020) was a term extraction challenge building on the ACTER dataset with data in:

- English, French, Dutch
- wind energy and corruption (training), dressage (equitation) (validation), heart failure (test)



Training	Test	Sequence Classifier			Token Classifier			NMT			Previous SOTA		
		Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
EN	EN	30.9	84.0	45.2	54.9	62.2	58.3	45.7	63.5	53.2	34.8	70.9	46.7
FR	EN	31.1	79.5	44.7	56.7	36.2	44.2	50.0	59.3	54.2			
NL	EN	22.3	91.1	35.9	55.3	61.8	**58.3**	48.3	64.3	55.2			
ALL	EN	31.4	85.8	*46.0*	54.4	58.2	56.2	50.2	61.6	*55.3*			
EN	FR	34.6	79.0	*48.1*	65.4	51.4	**57.6**	48.8	61.3	54.4			
FR	FR	32.2	80.2	46.0	68.7	43.0	52.9	52.7	59.6	55.9	44.2	51.5	48.1
NL	FR	26.1	84.7	40.0	62.3	48.5	54.5	54.3	60.9	57.4			
ALL	FR	33.2	78.9	46.7	62.7	49.4	55.3	55.0	60.4	*57.6*			
EN	NL	42.8	89.8	*58.0*	67.9	71.7	**69.8**	48.8	63.9	55.4			
FR	NL	41.3	87.6	56.1	69.2	55.2	61.4	56.2	63.4	59.6			
NL	NL	32.7	94.1	48.5	71.4	67.8	69.6	60.6	70.7	*65.2*	18.9	18.6	18.7
ALL	NL	40.4	91.5	56.0	70.0	65.8	67.8	60.6	70.0	64.9			

Relation Extraction


causalRelation(COVID-19, cough)
XLM-R
cough. COVID-19. The cough was caused by COVID-19

Prespecified relation typology:

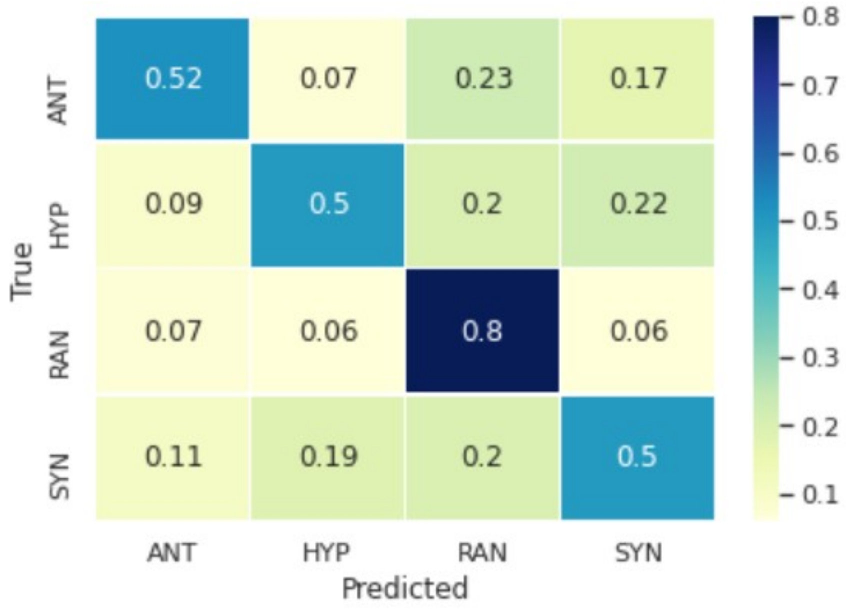
- activity relation (actor - activity, etc.)
- causal relation (cause - effect, etc.)
- generic relation (is_a)
- partitive relation (parts - whole)
- …

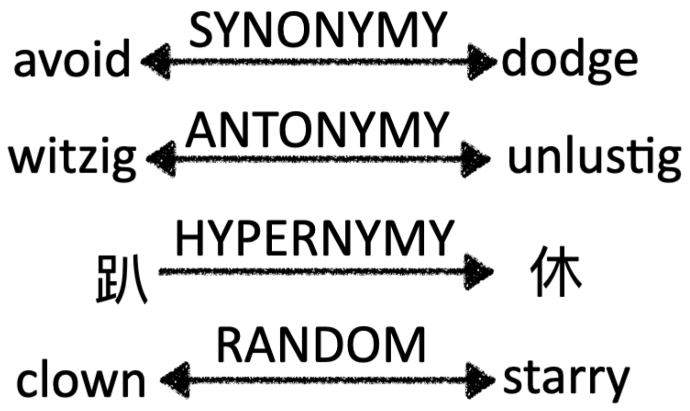
Evaluation in progress

- Adapting existing datasets to our typology:
  - SemEval 2007 Task 4: Classification of Semantic Relations between Nominals
  - SemEval 2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals
  - WCL hypernym dataset
- Generating new datasets for the task:
  - manually generating TSC from multilingual texts (two experts in German + students in other languages)
  - automatically annotating synonyms in sentences based on patterns, e.g. long form + acronym vs. acronyms without synonyms in sentences
  - manual annotation for negative examples, i.e., no relation to be predicted

Context-Free Relation Extraction

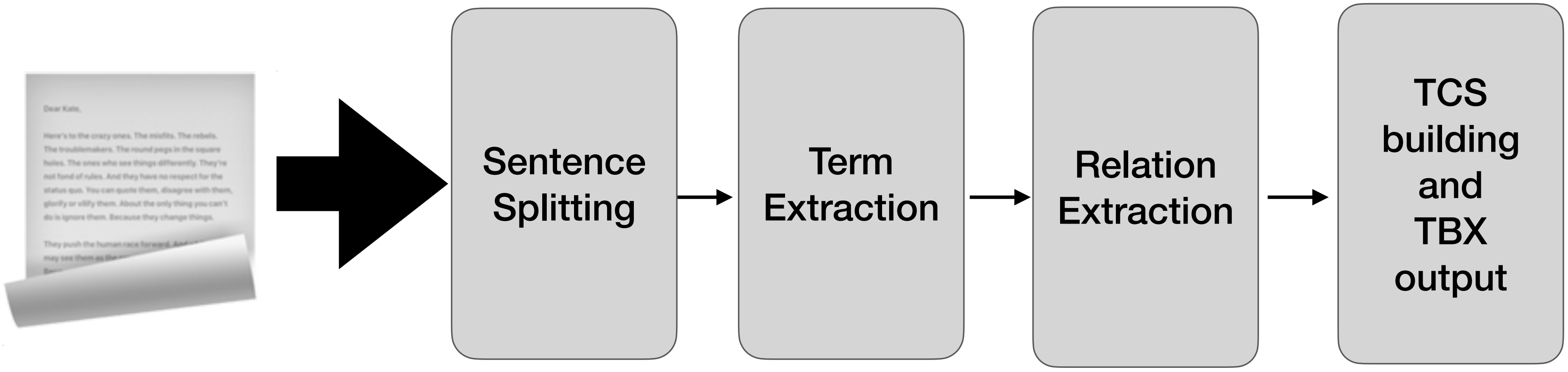
Winning system at the Cognitive Aspects of the Lexicon (CogALex) Shared Task 2020: 4 languages, 3 relations + random





Wachowiak, L., Lang, C., Heinisch, B. & Gromann, D. (2020) CogALex-VI Shared Task: Transrelation - A Robust Multilingual Language Model for Multilingual Relation Identification. CogALex VI Proceedings@COLING.

Sentence-Level TCS Learning

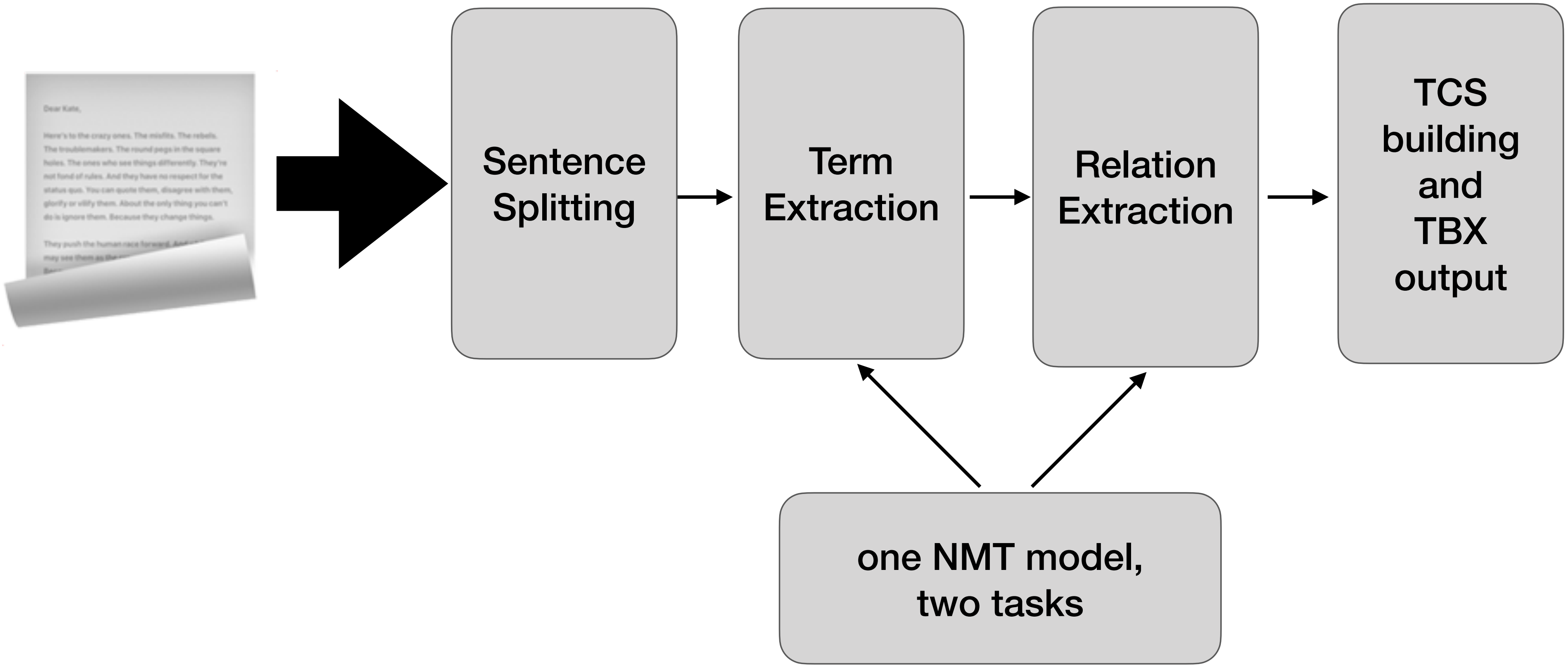


Dear Kate,

Here's to the crazy ones. The misfits. The rebels. The troublemakers. The round pegs in the square holes. The ones who see things differently. They're not fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things.

They push the human race forward. And
may see them as the

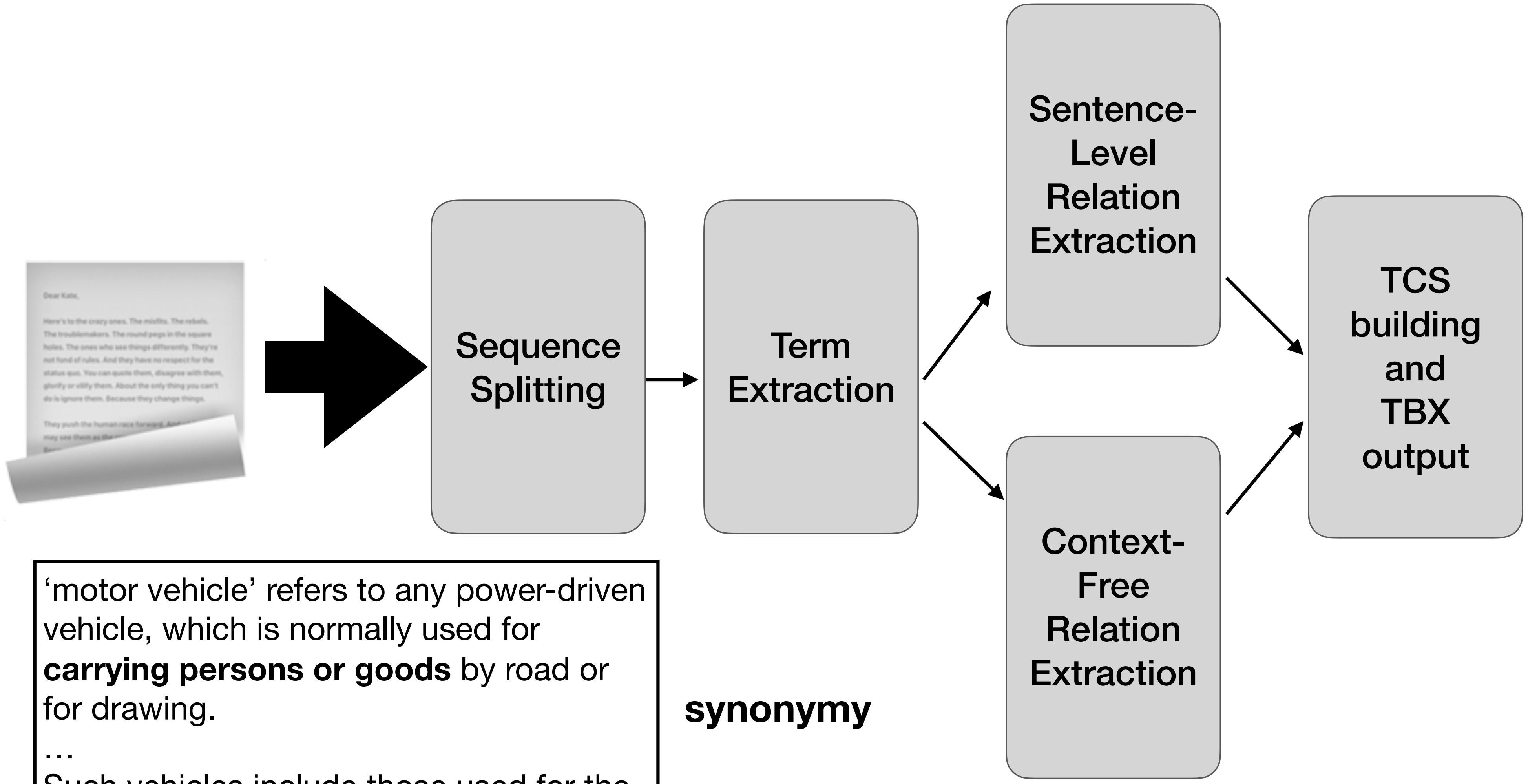
Sentence-Level TCS Learning



Sentence Splitting → Term Extraction → Relation Extraction → TCS building and TBX output

one NMT model, two tasks

Globalized Sentence-Level TCS Learning


Sequence Splitting
Term Extraction
Sentence-Level Relation Extraction
Context-Free Relation Extraction
TCS building and TBX output

‘motor vehicle’ refers to any power-driven vehicle, which is normally used for **carrying persons or goods** by road or for drawing.
…
Such vehicles include those used for the **carriage of persons or goods**.

synonymy

Text2TCS Team

DAGMAR GROMANN
Project leader

LENNART WACHOWIAK
Machine learning and IT

CHRISTIAN LANG
Translation and IT

BARBARA HEINISCH
Terminology and usability

Challenges and next steps

- very few training datasets for specific relation types, e.g. ownership or developmental relation, also for negative examples, i.e., no relation

- creating TCS data manually
  - time- and human resource-intensive
  - near L1 speakers for all languages to be included
  - domain expertise required on top

- Multilingual but not cross-lingual TCS learning
  - alignment across TCS in different languages
  - handling terminological gaps, e.g. *alunizaje (en: ram raid, de: ???), Schadenfreude (..:???), etc.*

Conclusion

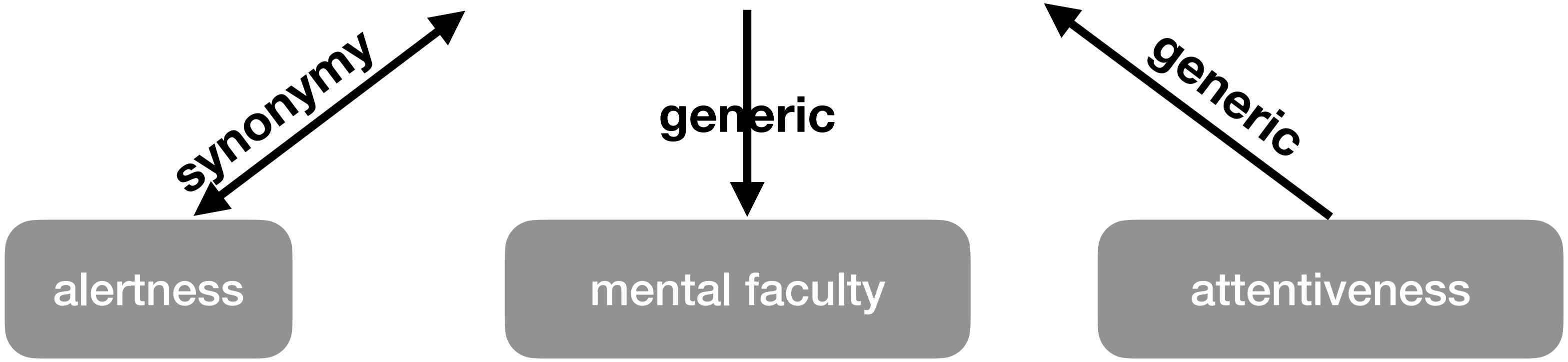
- structured and high-quality terminologies substantially contribute to specialized multilingual communication as well as translation, localization, etc.

- pretrained multilingual language models are highly performant on term extraction and relation extraction

- joint sentence-level extraction of terms, grouping them to synonyms, and learning their interrelations is still challenging

- Where to go from here?

Thank you for your attention!



References

Budin, Gerhard (1996). Wissensorganisation und Terminologie: Die Komplexität und Dynamik Wissenschaftlicher Informations- und Kommunikationsprozesse. Hartwig Kalverkämper (ed), Forum für Fachsprachen-Forschung, 28, Narr: Tübingen.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. ACL 2020.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Drewer, Petra & Schmitz, Klaus-Dirk (2017) *Terminologiemanagement: Grundlagen-Methoden-Werkzeuge*. Springer-Verlag.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726-742.

Petrucci, G., Rospocher, M., & Ghidini, C. (2018). Expressive ontology learning as neural machine translation. *Journal of Web Semantics*, *52*, 66-82

Rigouts Terryn, A., Hoste, V., Drouin, P., & Lefever, E. (2020). Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)* (pp. 85-94). European Language Resources Association (ELRA).

Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., ... & Sun, M. (2019). DocRED: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.