From Research to Production: Fine-Grained Analysis of Terminology Integration

Toms Bergmanis*, Mārcis Pinnis*, Paula Reichenberg**

* Tilde, Vienības gatve 75A, Riga, Latvia, LV-1004

** Hieronymus, Stauffacherstrasse 100 CH-8004 Zürich, Switzerland

TILDE

HIERONYMUS

Translations by Lawyers for Lawyers

Thesis

Terminology integration is a **cascade** of

1. terminology management
2. terminology identification
3. terminology translation

thus it is **prone** to problems due to **error propagation**.

Photo credit: https://www.watergardeningdirect.com/acatalog/Neptune-Blue-Ceramic-Solar-Cascade-Water-Feature.html

Outline

1. Three aspects of Terminology Integration:
    - Terminology Management
    - Terminology Identification
    - Terminology Translation
2. Main takeaways

Terminology Management

- Terminology for humans is not the same as terminology for machines
- Humans can:
  - Disambiguate based on external/world knowledge and experience
  - Work with corrupted/noisy data
- How do we get to terminology that is useful for machines?

Terminology Management

Common issues:

- Specificity
  - ✗ sport, prize, China
    (Source: IATE, Dinu et.al 2019)
  - ✗ deaths, transmission, close contact, face mask
    (Source: WMT 2021 Terminology task)
  - ✓ angular ball bearing, ball peen hammer, companion flange
    (Source: Bergmanis and Pinnis 2021)

Solution: use Inverse Document Frequency based filtering of your glossary!

Terminology Management

Common issues:

- Specificity
- Ambiguity
  - ✕ sense ambiguity: *organ*

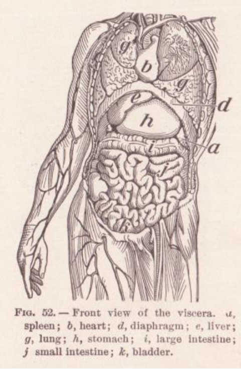
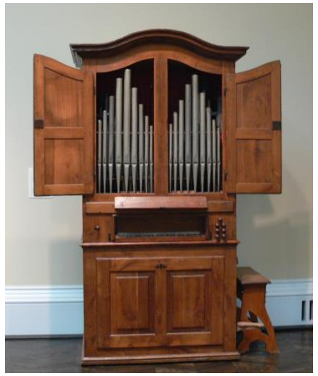
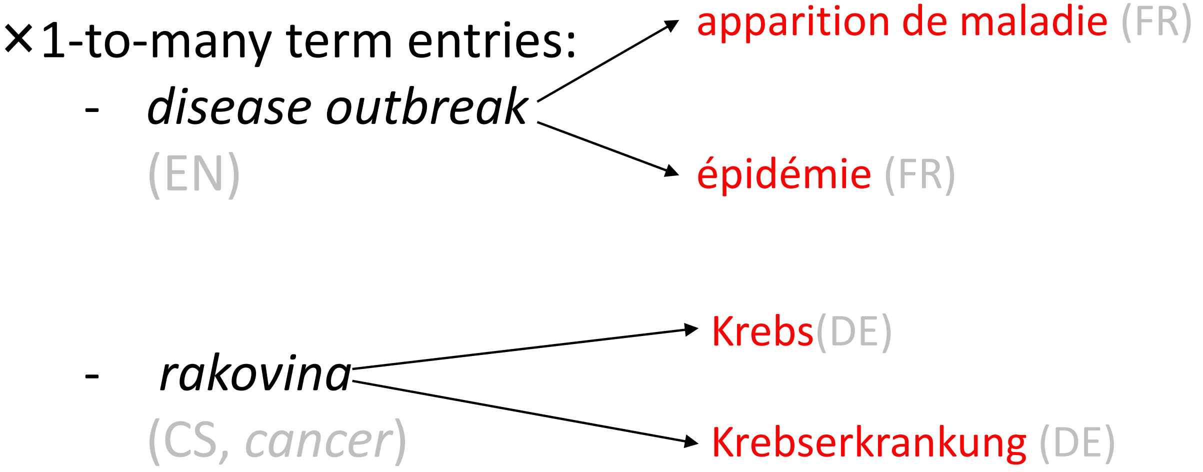


FIG. 52.— Front view of the viscera. a, spleen; b, heart; d, diaphragm; e, liver; g, lung; h, stomach; i, large intestine; j small intestine; k, bladder.





✕ 1-to-many term entries:

- *disease outbreak* (EN) → apparition de maladie (FR), épidémie (FR)
- *rakovina* (CS, *cancer*) → Krebs(DE), Krebserkrankung (DE)

(Source: WMT 2021 Terminology task)

Solution: filter ambiguous terms and commit to just one translation per collection!

Terminology Management

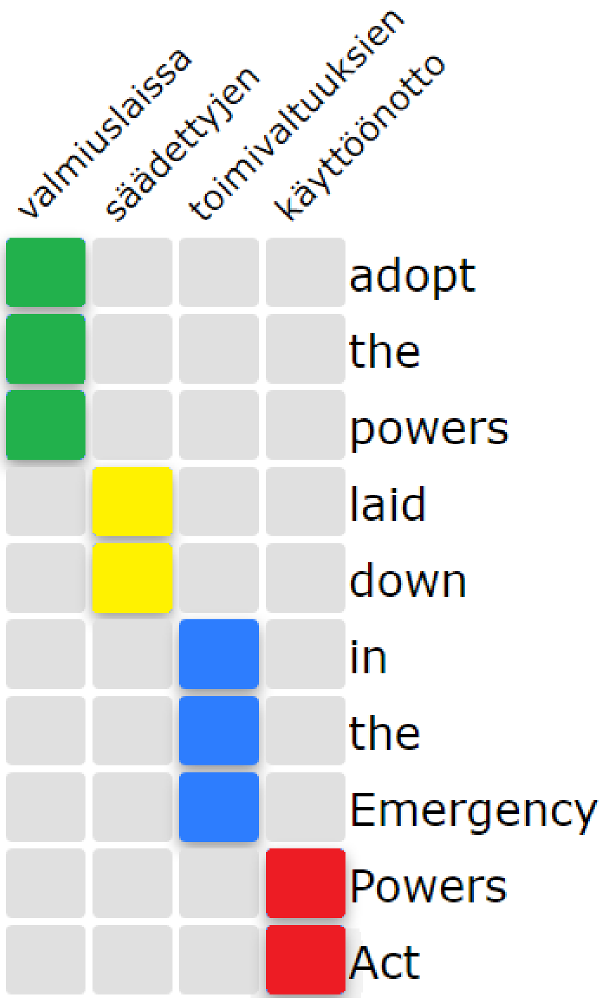
Common issues:

- Specificity
- Ambiguity
- Needless wordiness:

  ✕ adopt the powers laid down in the Emergency Powers Act

  =

  valmiuslaissa säädettyjen toimivaltuuksien käyttöönotto



valmiuslaissa	säädettyjen	toimivaltuuksien	käyttöönotto	
🟩				adopt
🟩				the
🟩				powers
	🟨			laid
	🟨			down
		🟦		in
		🟦		the
		🟦		Emergency
			🟥	Powers
			🟥	Act

https://nlg.isi.edu/demos/picaro/

Solution: decompose long multiword expressions when possible!

Terminology Management:
Type of terminological data

- **The minimalist's point of view** - a collection of bilingual term pairs for every domain
- **The maximalist's point of view** - a collection of bilingual term pairs with all the necessary meta-data:
  - Morphological information
  - Syntactic information
  - Domain information
- *The overwhelming majority of term collections used in practice are minimalist's term collections*

Terminology Identification

Common challenges:

- Morphological complexity
- Part-of-speech ambiguity*
- Term sense ambiguity*

* if unresolved using Terminology Management

Terminology Identification: Morphological Complexity

	Sing	Plural
NOM	vācietis	vācieši
GEN	vācieša	vāciešu
DAT	vācietim	vāciešiem
ACC	vācieti	vāciešus
INST	ar vācieti	ar vāciešiem
LOC	vācietī	vāciešos
VOC	vācieti!	vācieši!

- In morphologically complex languages terms can take many forms which hinder term identification
- Solution: use stemmer (fast, lower precision)
- Solution: use lemmatizer (slower, higher precision)

Latvian: vācietis (English: *a German*)

Terminology Identification: Part-of-speech ambiguity

Use the control. Control the execution.

A noun or a verb?

Dry clothes

A noun or an adjective?

This is clearly too ambiguous to tell

- Solution (partial): use morpho-syntactic taggers
- What if the term collection does not provide any morphological metadata?
  - Try enriching term collections automatically
  - Filter out terms that cannot be reliably supported

Terminology identification: Summary

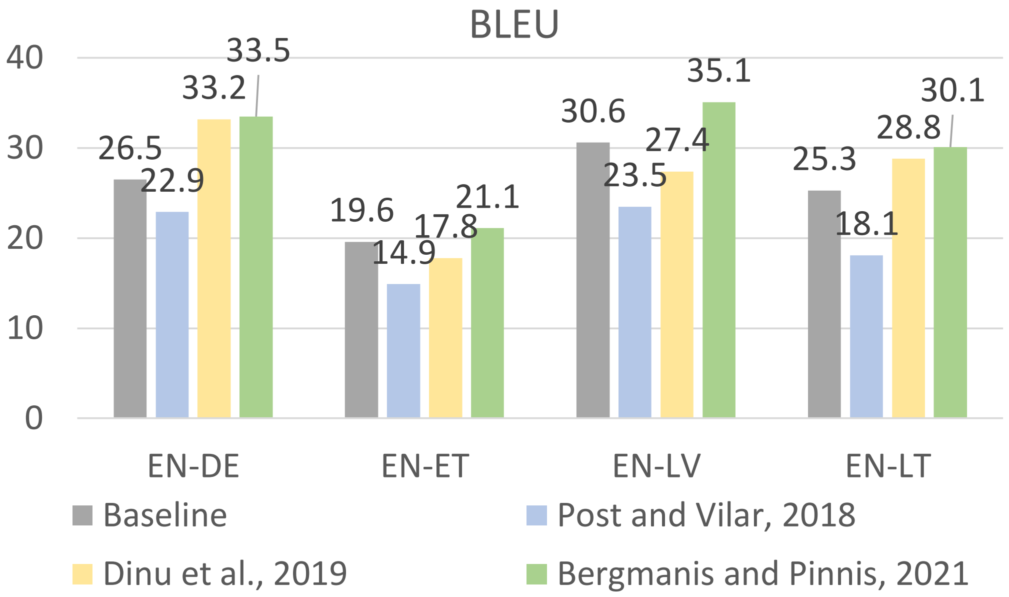
- A practical solution:
  - Filter term collections to not include:
    - General language
    - Ambiguous terms that cannot be reliably supported by your method
  - Then, if term collections are minimalistic:
    - depending on language and tools that are available, identify terms using either:
      - Lemmatization, or
      - stemming
  - *If term collections are meta-data-rich, let us know – we would like to see that with our own eyes*.

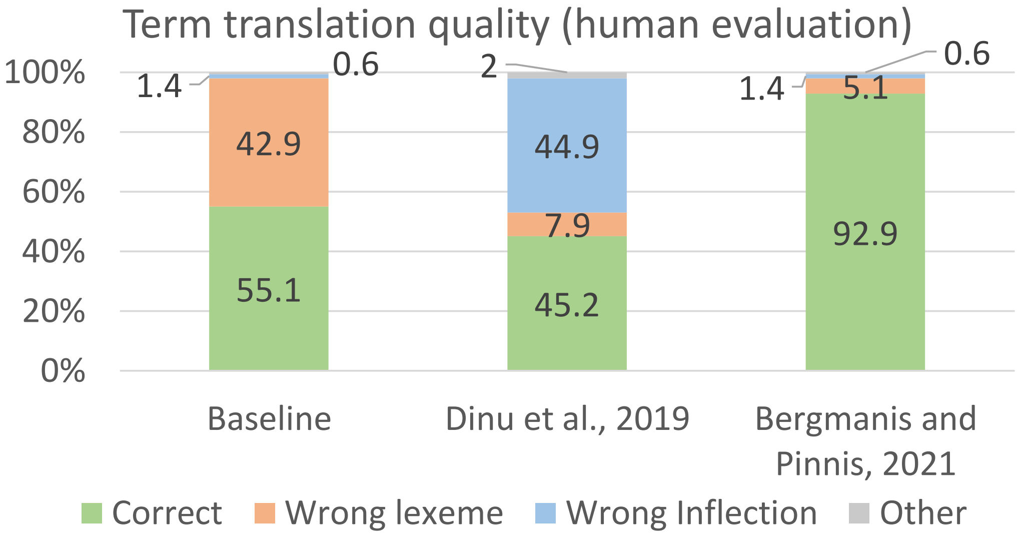
Terminology Translation

- When we have a term collection and we can identify terms in the source text, what are our integration options?
  - Constrained Decoding (Post and Vilar, 2018)
  - Exact Target Annotations (Dinu et al., 2019)
  - Target Lemma Annotations (TLA) (Bergmanis and Pinnis, 2021)

Terminology Translation

- We use Target Lemma Annotations since they allow achieving the highest overall translation quality and term translation accuracy for morphologically rich languages
- For languages with simple nominal morphology, other methods (Post and Vilar, 2018; Dinu et al. 2019) are also viable





*Results from Bergmanis and Pinnis, 2021

Terminology Translation: Target Lemma Annotation

Latvian (Target): *Rīks , kas der uzgriežņa galvai .*

Latvian Lemmas: *Rīks , kas derēt uzgrieznis galva .*

Word Alignments: *0-1 2-2 3-3 4-8 5-5 6-9*

English (Source): *A tool that fits the head of the nut .*

English with TLA: A tool that <fits|derēt> the head of the <nut|uzgrieznis>

We use linguistic input features (Sennrich and Haddow 2016) to facilitate annotation on the source side

* Example from Bergmanis and Pinnis, 2021

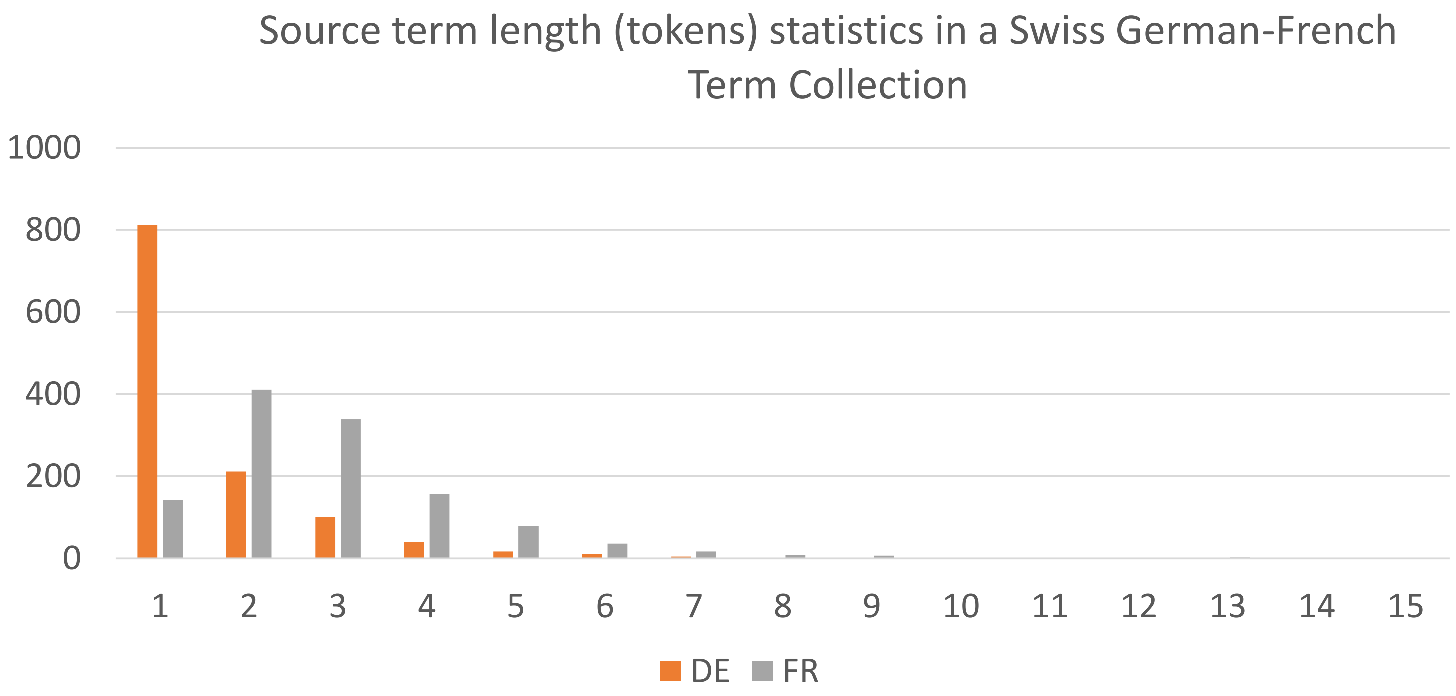
Terminology Translation: From Research to Production

- The goal of research – to publish
- The goal of production – to deliver a reliable product

- The main question that arose when deploying terminology integration in production:
  - How to prepare training data such that the trained systems will be capable of handling terms used by customers?

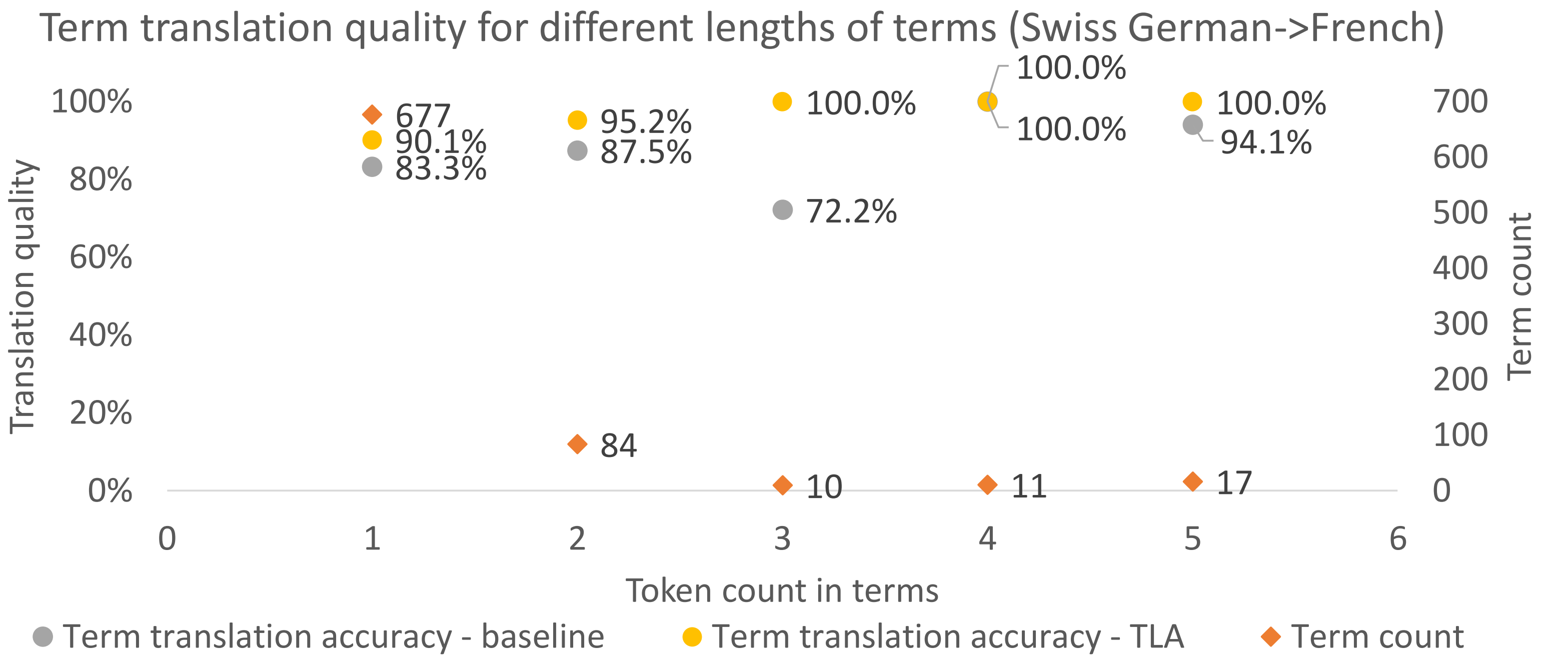
Terminology Translation: From Research to Production

- Challenge - Term length



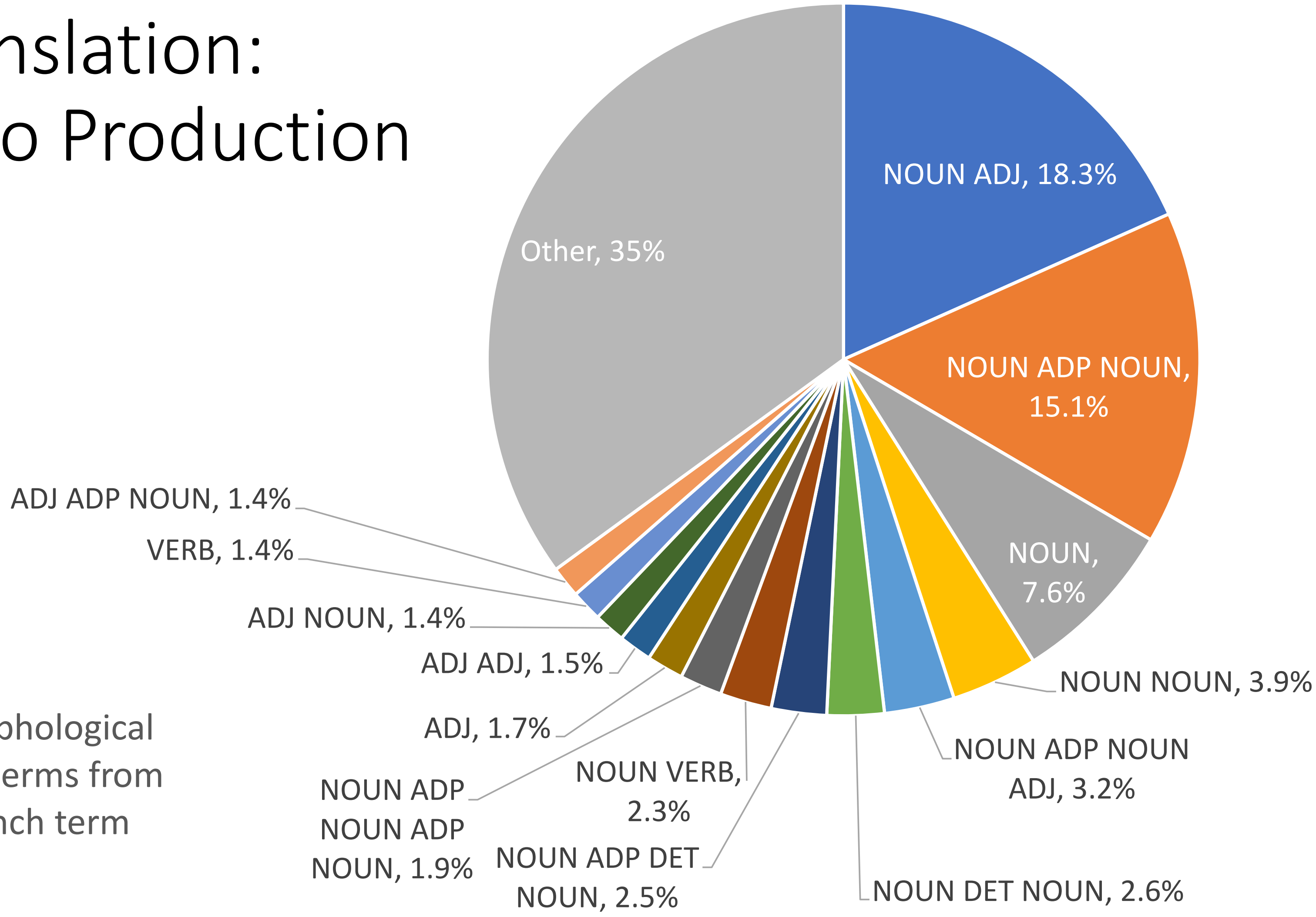
Terminology Translation: From Research to Production

- Challenge - Term length
- Solution – annotate multi-word phrases with TLA



Terminology Translation: From Research to Production

- Challenge – multiword terms have complex syntactic structure

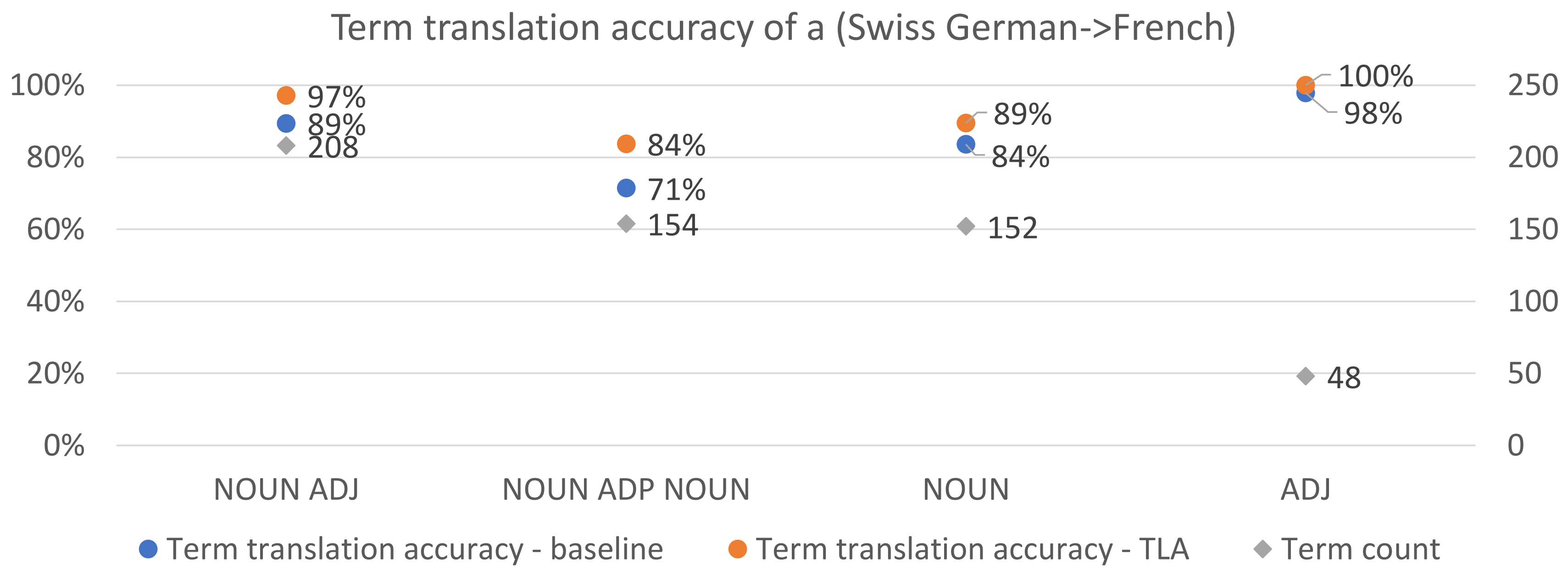


Statistics of the morphological structure of French terms from a Swiss German-French term collection

** Note that the part of speech tags were acquired using an automatic part-of-speech tagger and may be noisy!*

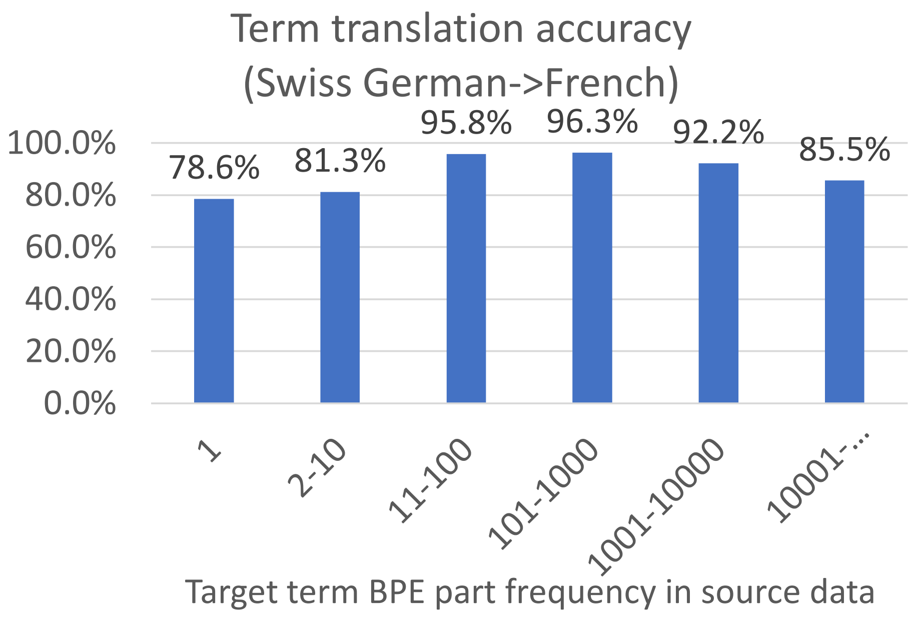
Terminology Translation:
From Research to Production

- Challenge – multiword terms have complex syntactic structure
- Solution – make sure that you annotate phrases with syntactic structures representing terms



Terminology Translation: From Research to Production

- Challenge – some terms consist of rare BPE parts and are translated poorly
- Solution 1 – make sure that training data TLA contain BPE parts relevant to terms used at the test time
- Solution 2 – filter term collections such that out-of-vocabulary terms are ignored
- Solution 3 – use character representations of TLA (Niehues, 2021)



Main Takeaway

- Terminology integration is a **cascade** of terminology creation, curation, identification and only then translation using MT.
- Terminology creation and curation is and should be done by **professional translators and domain experts**.
- **Poor terminology management choices will be propagated in downstream processes** – terminology identification and terminology translation, and will impede the final translation quality.

Main Takeaway

To mitigate error propagation, pay attention to how terminology is managed and prepared for MT such that it is MT-ready

- Make sure that terminology is consistent
- Make sure that terminology is domain-specific
- Do not overexaggerate with needless wordiness
    - Online/dynamic learning, and translation memories may be better suited for such data
- Provide enough metadata such that your term identification method is able to function properly

References

Sennrich, Rico, and Barry Haddow. "Linguistic Input Features Improve Neural Machine Translation." *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*. 2016.

Post, Matt, and David Vilar. "Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.

Dinu, Georgiana, et al. "Training Neural Machine Translation to Apply Terminology Constraints." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

Bergmanis, Toms, and Mārcis Pinnis. "Facilitating Terminology Translation with Target Lemma Annotations." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Niehues, Jan. "Continuous Learning in Neural Machine Translation using Bilingual Dictionaries." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.