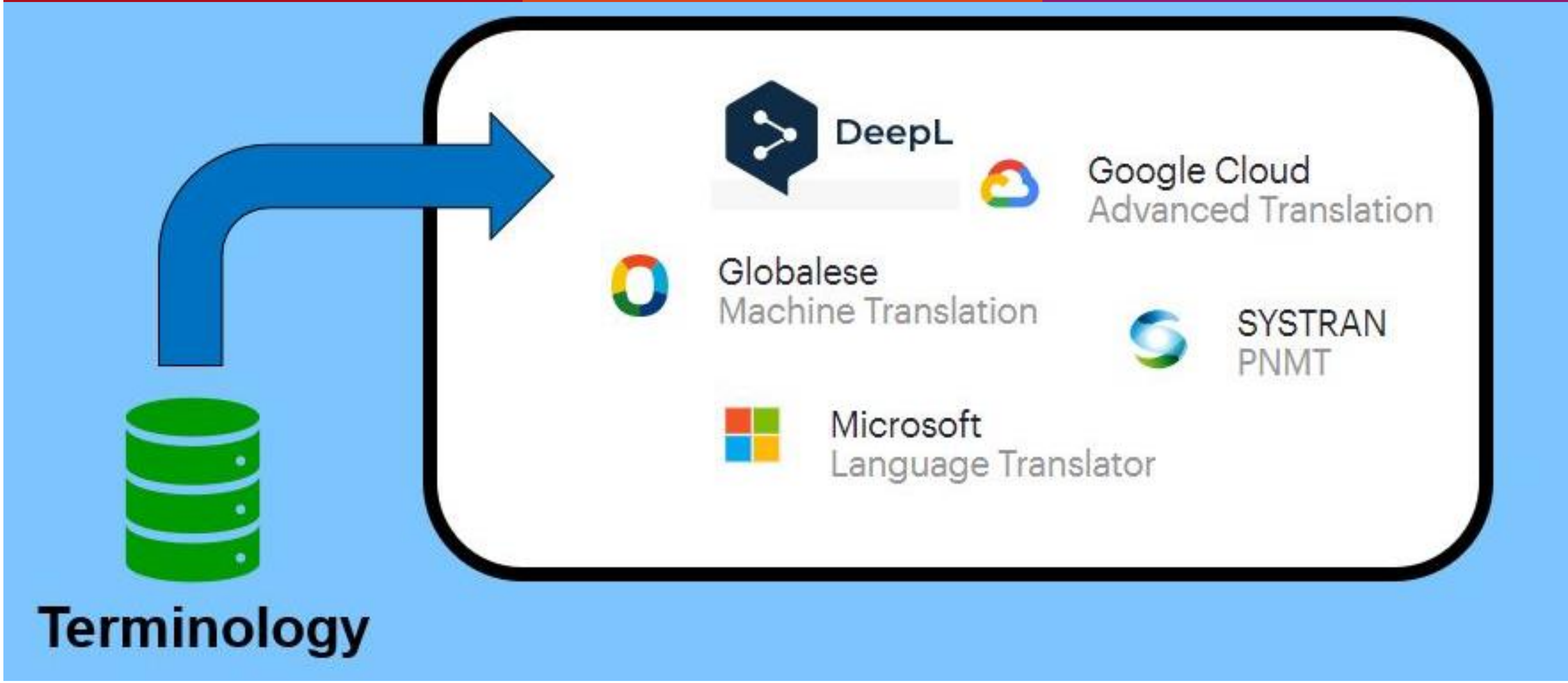


Integrating Terminology into NMT – Glossaries

Agenda

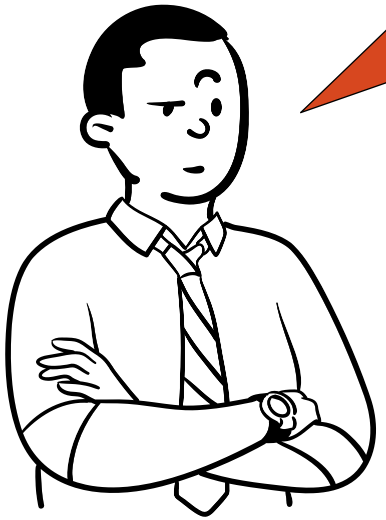
1) Quick paper recap
2) Terminology check during post-editing
3) Why Glossaries?
    a) Reducing post-editing effort
    b) Issue 1: Concept vs. term orientation
    c) Issue 2: How does the actual term integration work?
4) Demonstration
    a) From termbase to glossary
    b) DeepL and its glossary feature
5) Conclusion and outlook
6) References

1) Quick paper recap…

- Terminology integration can be part of
    - customising or domain-adapting a generic NMT engine
    - by re-training it with translation data from a specific domain
        - e.g. client or project TMs with consistent terminology

- <u>Different approaches</u>:
    - architecture-centric
    - data-centric
        - at decoding
        - **at training** (masking, data augmentation)

2) Terminology check during post-editing



Can‘t we just post-edit
the MT output?

We can always use a
termbase to check
the use of terminology...

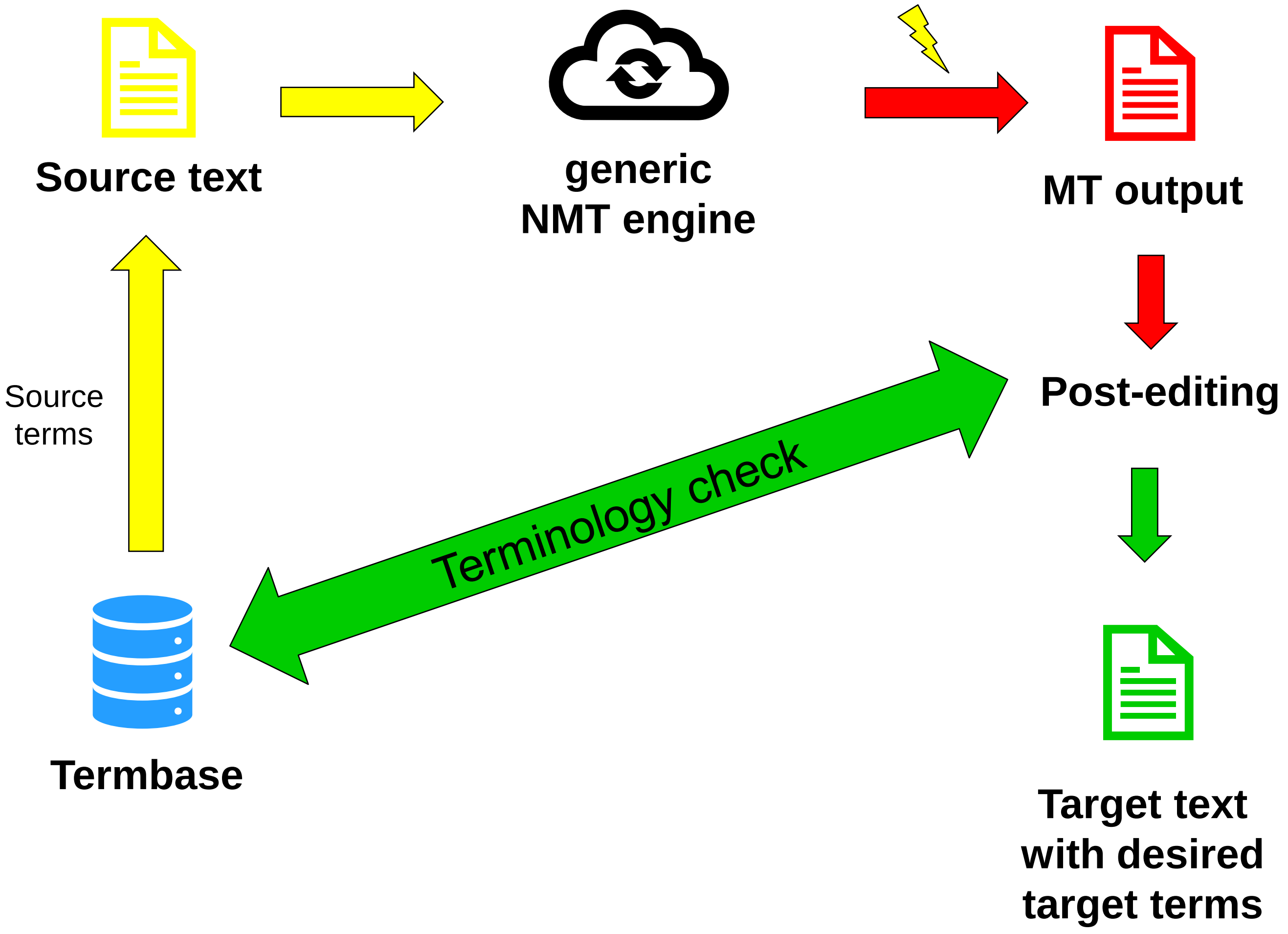
2) Terminology check during post-editing

- Ensuring consistent terminology in the target text forms part of *full post-editing* (ISO 18587:2017)

- Generic MT engines often struggle to produce output with consistent terminology due to:
  - synonyms/variants in the training data → *translation unit* vs. *source/target pair*
  - highly specific terms missing from the training data
  - complex term structure (e.g., compounds and multi-word terms) → high-efficiency particulate air filter

- Such an MT output may include:
  - mistranslations of homographs and acronyms (depending on the amount of context!)
    → *mole* = *Muttermal* (*birthmark*) vs. *Maulwurf* (*animal*)
    → *NMT* = *NMT* vs. *NMÜ* (often copied from the source text)
  - omissions of whole terms or word parts
  - inconsistent terminology (e.g., Winter/Zielinski 2020:216–221)

2) Terminology check during post-editing



- detecting missing or forbidden target terms in the MT output

- Terminology check with QA component in a CAT tool or with a stand-alone QA tool such as QA Distiller
(if you want to learn more, see our advanced learning nugget on that topic)

2) Terminology check during post-editing

Efficient (full) post-editing of NMT output requires:

- a well-maintained termbase
  - if possible, avoid **amitted** synonyms and variants and (Seidel/Grützmacher 2020:188f.)
  - include forbidden terms (both from text producers and the MT engine)
  - if you use homographs and acronyms, manage them with clear data categories or cross-references
- Competent post-editors (Nitzke et al. 2019)
  - bilingual competency
  - **AND** extralinguistic competency (domain knowledge), among others

2) Terminology check during post-editing

Well, post-editing all terminology mistakes does sound like a lot of work!

Any way to reduce this effort?

3a) Why Glossaries? Reducing post-editing effort

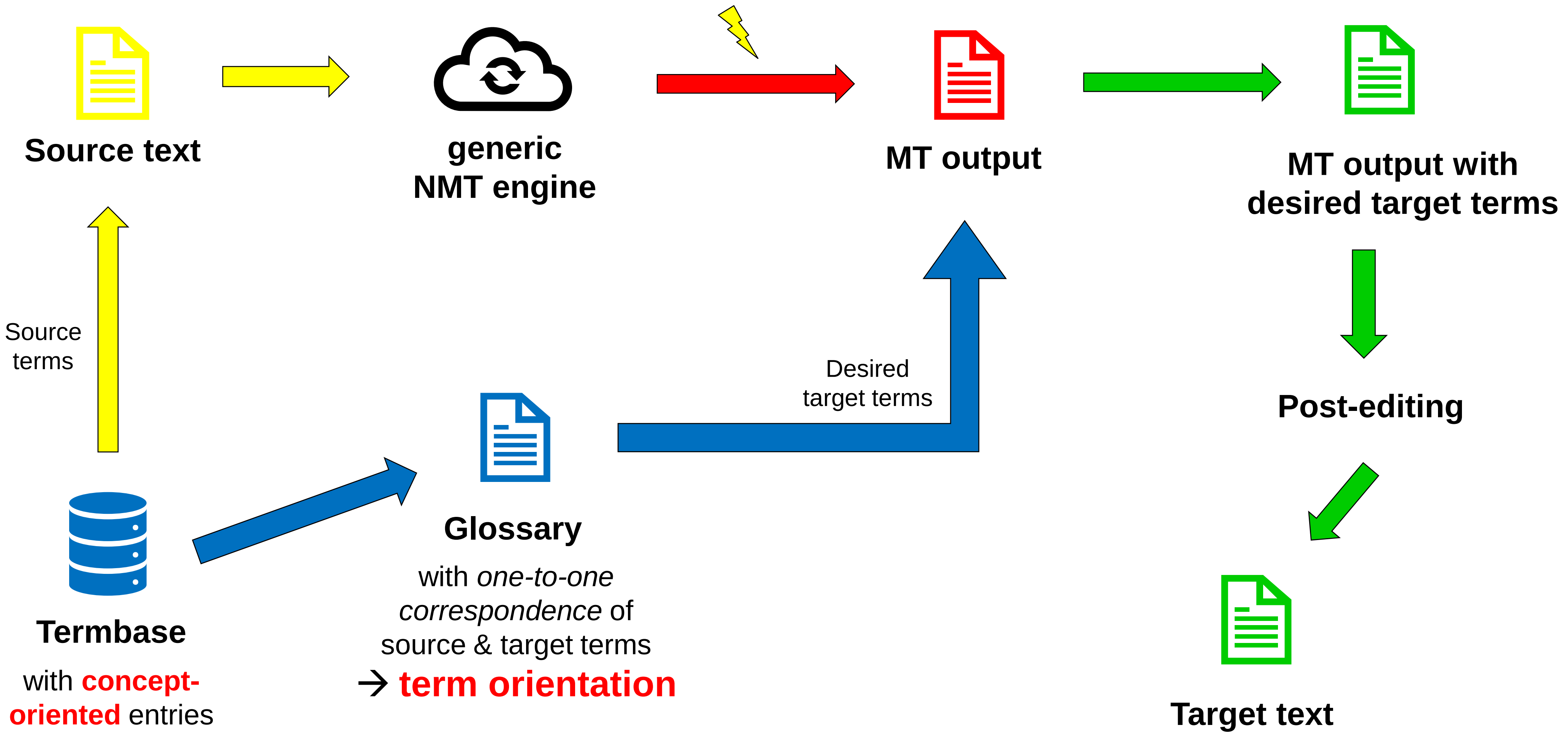
- Besides customising or domain-adapting a generic NMT engine, you can use ***glossaries*****.**

- What is a glossary?
  - “terminological dictionary
  - that contains designations [terms]
  - from one or more domains or subjects
  - together with equivalents in one or more natural languages” (ISO 1087 2019:16)

English	German
ant	Ameise
butterfly	Schmetterling
mole	Maulwurf

- Several NMT providers/systems offer custom terminology support via glossaries
  - e.g., DeepL, IBM Watson, SYSTRAN, Globalese, Google Cloud, among others
  - numbers rising → seven in 2021, nine in 2022 (Intento 2021, 2022)

3a) Why Glossaries? Reducing post-editing effort



Source text

generic NMT engine

MT output

MT output with desired target terms

Source terms

Desired target terms

Post-editing

Glossary

with *one-to-one correspondence* of source & target terms

→ **term orientation**

Termbase

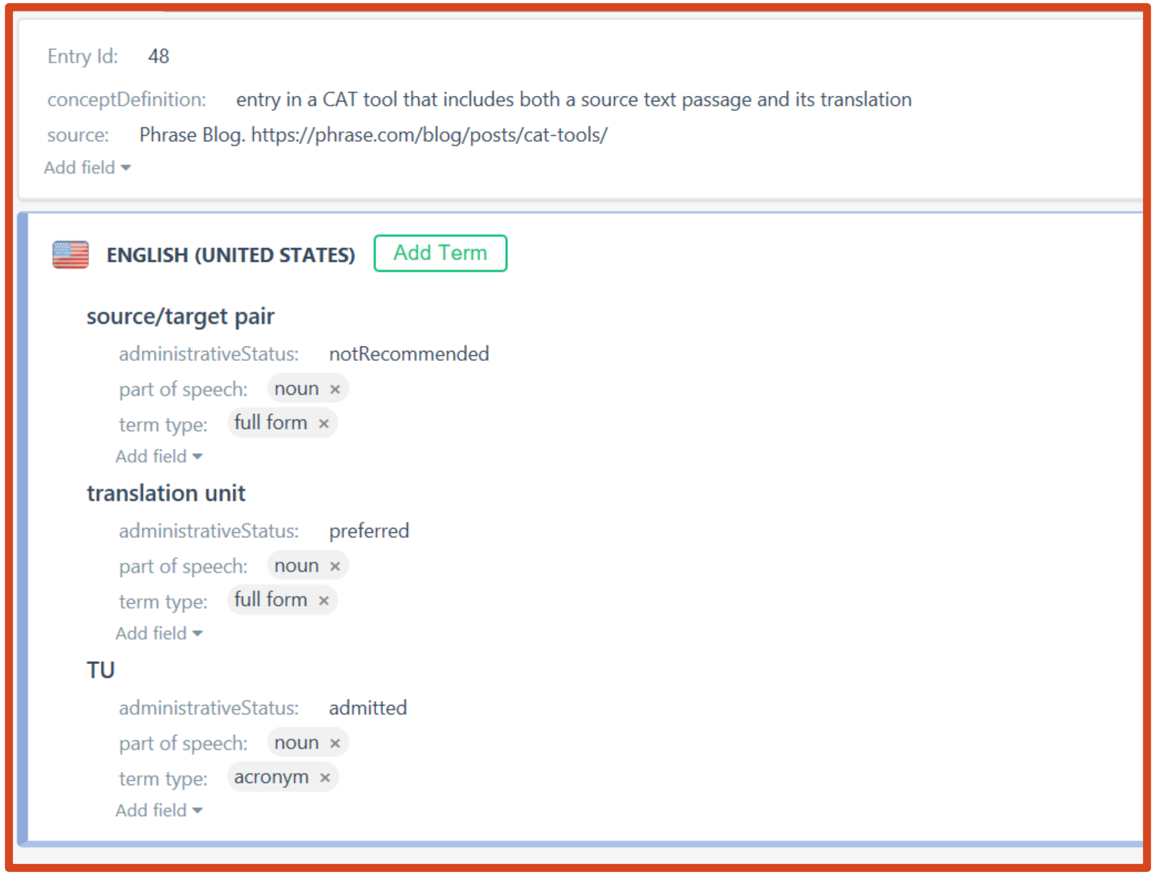
with **concept-oriented** entries

Target text

3b) Issue 1: Concept vs. term orientation

Termbases are concept-oriented:
(e.g., ISO 12616-1:2021)

- 1 concept = 1 entry
- All terms (incl. synonyms, abbreviations, variants) and all relevant metadata
- e.g. *administrative status* to perform terminology checks in writing and translation
  - *preferred*, *admitted*, *notRecommended*, *obsolete* (DataCatInfo: 'administrative status' in TermWeb)



Entry Id: 48
conceptDefinition: entry in a CAT tool that includes both a source text passage and its translation
source: Phrase Blog. https://phrase.com/blog/posts/cat-tools/

ENGLISH (UNITED STATES)

source/target pair
administrativeStatus: notRecommended
part of speech: noun ×
term type: full form ×

translation unit
administrativeStatus: preferred
part of speech: noun ×
term type: full form ×

TU
administrativeStatus: admitted
part of speech: noun ×
term type: acronym ×

3b) Issue 1: Concept vs. term orientation

- In many cases, NMT engines can’t yet use concept-oriented terminology data
- need for **glossaries** (term lists) → e.g., in XLSX, CSV, TSV

- **one-to-one term correspondence** (Winter 2021:6–8)

1) all source language terms (even forbidden ones) → complete check
2) ONE preferred target language equivalent (respectively)
3) no ambiguous source language terms
   to avoid substitutions with incorrect target language terms*

This could possibly exclude relevant terms from being integrated in the MT output!

consideration: Does the alternative meaning of the term occur in the source texts?

English	German
mole	Maulwurf (animal)
mole	Muttermal (*birthmark*)
mole	…

3c) Issue 2: How does the actual term integration work?

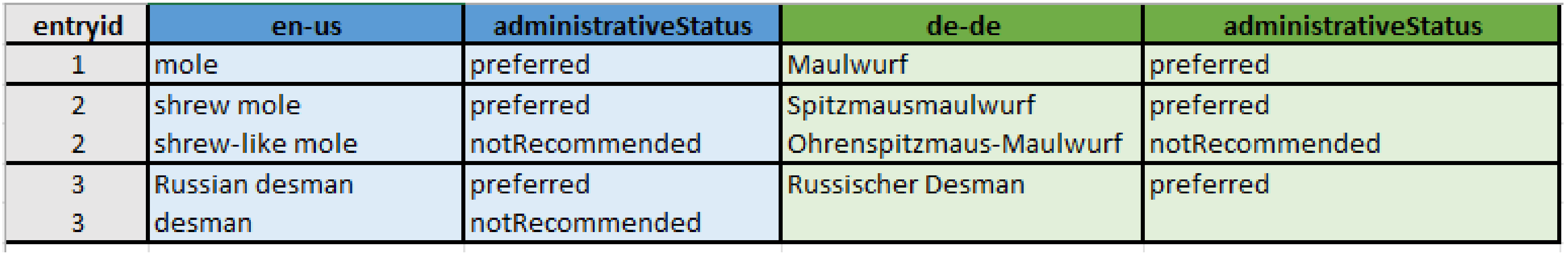
- Term substitution is often only based on string matching, i.e. *Find & Replace*

- If your glossary doesn't contain inflected forms (e.g., conjugated verbs, declined nouns),
  - Only lemmas are found and replaced → limited use (proper names, slogans, etc.)
  - Data augmentation at least manages inflected forms more or less reliably
    (but only when the source term is detected) (Winter 2021:6–8)

- Solution: (Winter 2021:6–8)
  1) Include inflected forms in your glossary to improve term detection (overkill?)
  2) Morphosyntactic terminology integration (as provided by e.g. DeepL)
     - adapts integrated terms and their context (articles, adjectives etc.)
       to the grammar of the target language (e.g. grammatical gender/number)

4a) Demonstration: from termbase to glossary

Concept-oriented terminology entries

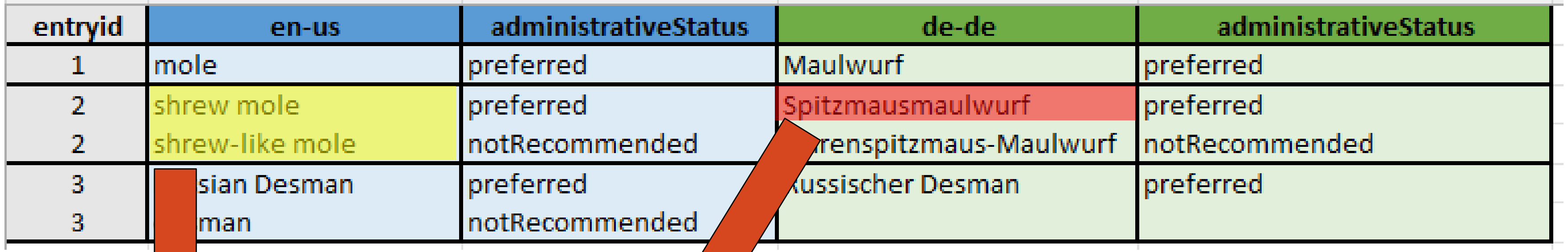


entryid	en-us	administrativeStatus	de-de	administrativeStatus
1	mole	preferred	Maulwurf	preferred
2	shrew mole	preferred	Spitzmausmaulwurf	preferred
2	shrew-like mole	notRecommended	Ohrenspitzmaus-Maulwurf	notRecommended
3	Russian desman	preferred	Russischer Desman	preferred
3	desman	notRecommended		

But how do we get a glossary now?

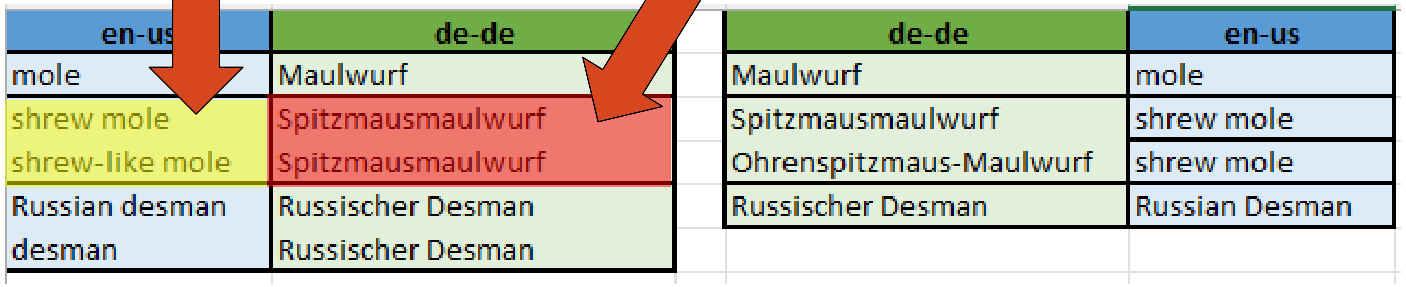
4a) Demonstration: from termbase to glossary

Concept-oriented terminology entries



entryid	en-us	administrativeStatus	de-de	administrativeStatus
1	mole	preferred	Maulwurf	preferred
2	shrew mole	preferred	Spitzmausmaulwurf	preferred
2	shrew-like mole	notRecommended	renspitzmaus-Maulwurf	notRecommended
3	sian Desman	preferred	ussischer Desman	preferred
3	man	notRecommended		

Term-oriented glossary entries → two lists, depending on the language pair



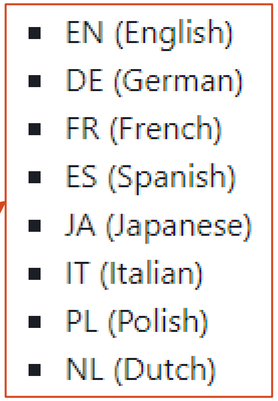
en-us	de-de
mole	Maulwurf
shrew mole	Spitzmausmaulwurf
shrew-like mole	Spitzmausmaulwurf
Russian desman	Russischer Desman
desman	Russischer Desman

de-de	en-us
Maulwurf	mole
Spitzmausmaulwurf	shrew mole
Ohrenspitzmaus-Maulwurf	shrew mole
Russischer Desman	Russian Desman

4b) Demonstration: DeepL and its glossary feature

- MT output produced by DeepL (i.e., by a generic NMT engine)
  - any of the terminology mistakes discussed earlier?

- How does the glossary feature work?
  - today → free desktop version: 1 glossary, up to 10 entries
  - glossary support
    - for web and desktop (since May 2020) (DeepL 2020)
    - for the API (since August 2021) (DeepL 2021)
    - in the formats CSV/TSV and for 28 language pairs (DeepL n. d.)



- EN (English)
- DE (German)
- FR (French)
- ES (Spanish)
- JA (Japanese)
- IT (Italian)
- PL (Polish)
- NL (Dutch)

4b) Demonstration: DeepL and its glossary feature

Source text sample:

“[Moles] are small, dark-furred animals with cylindrical bodies and hairless, tubular snouts.

They range in size from the tiny shrew moles of North America, as small as 10 cm in length and weighing under 12 grams, to the Russian desman, with a body length of 18–22 cm, and a weight of about 550 grams.

The fur varies between species, but is always dense and short; desmans have waterproof undercoats and oily guard hairs, while the subterranean moles have short, velvety fur lacking any guard hairs.” (Wikipedia 2023, ‘Talpidae’)

Original MT output:

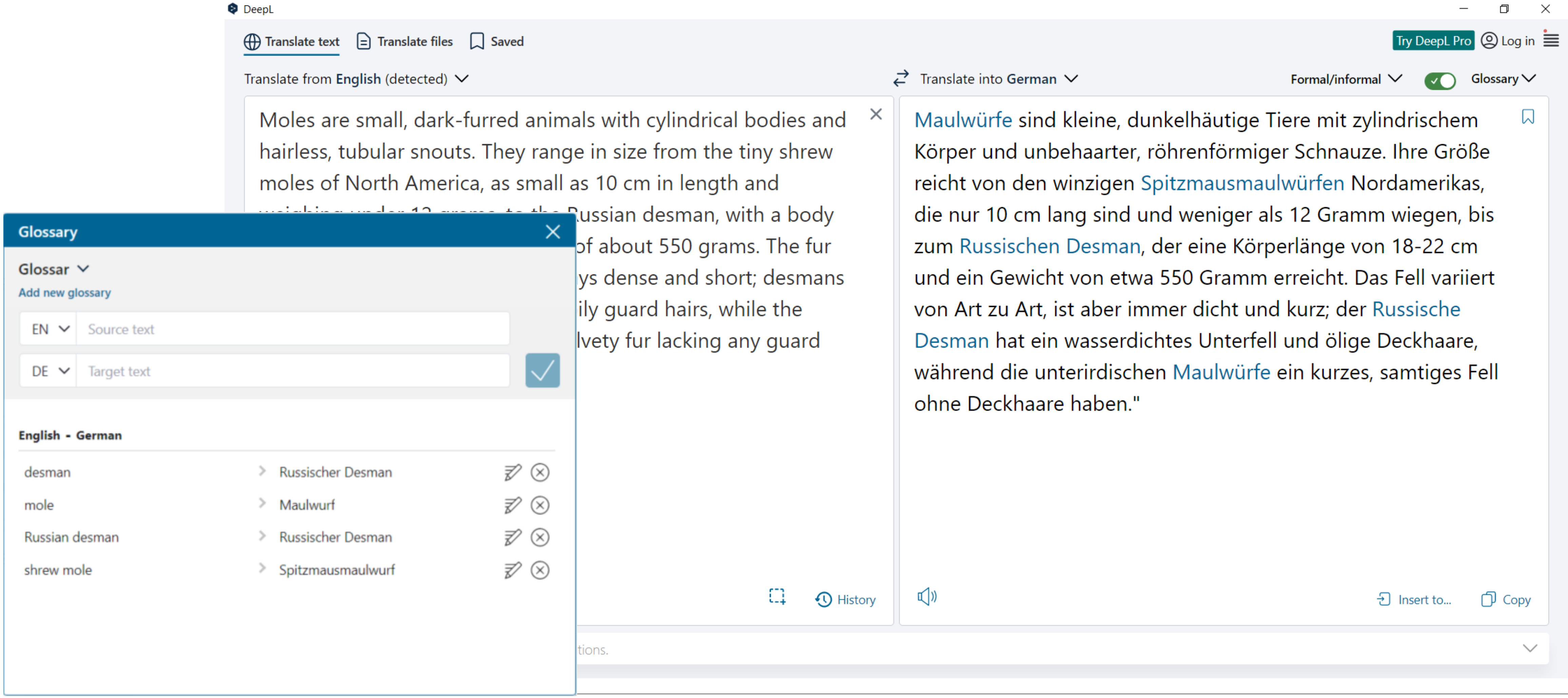
„Maulwürfe ✓ sind kleine, dunkelhäutige Tiere mit zylindrischen Körpern und unbehaarten, röhrenförmigen Schnauzen.

Ihre Größe reicht von den winzigen nordamerikanischen Spitzmäusen ✗, die nur 10 cm lang sind und weniger ○ als 12 Gramm wiegen, bis zum russischen Desman mit einer Körperlänge von 18-22 cm und einem Gewicht von etwa 550 Gramm.

Das Fell variiert von Art zu Art, ist aber immer dicht und kurz; Desmans ✗ haben ein wasserdichtes Unterfell und ölige Deckhaare, während die unterirdischen Maulwürfe ✓ ein kurzes, samtiges Fell ohne Deckhaare haben.“

(DeepL, 20 February 2023)

4b) Demonstration: DeepL and its glossary feature (see video)



Moles are small, dark-furred animals with cylindrical bodies and hairless, tubular snouts. They range in size from the tiny shrew moles of North America, as small as 10 cm in length and

Russian desman, with a body

of about 550 grams. The fur

ys dense and short; desmans

ily guard hairs, while the

lvety fur lacking any guard

Maulwürfe sind kleine, dunkelhäutige Tiere mit zylindrischem Körper und unbehaarter, röhrenförmiger Schnauze. Ihre Größe reicht von den winzigen Spitzmausmaulwürfen Nordamerikas, die nur 10 cm lang sind und weniger als 12 Gramm wiegen, bis zum Russischen Desman, der eine Körperlänge von 18-22 cm und ein Gewicht von etwa 550 Gramm erreicht. Das Fell variiert von Art zu Art, ist aber immer dicht und kurz; der Russische Desman hat ein wasserdichtes Unterfell und ölige Deckhaare, während die unterirdischen Maulwürfe ein kurzes, samtiges Fell ohne Deckhaare haben."

Glossary

Glossar

Add new glossary

EN Source text

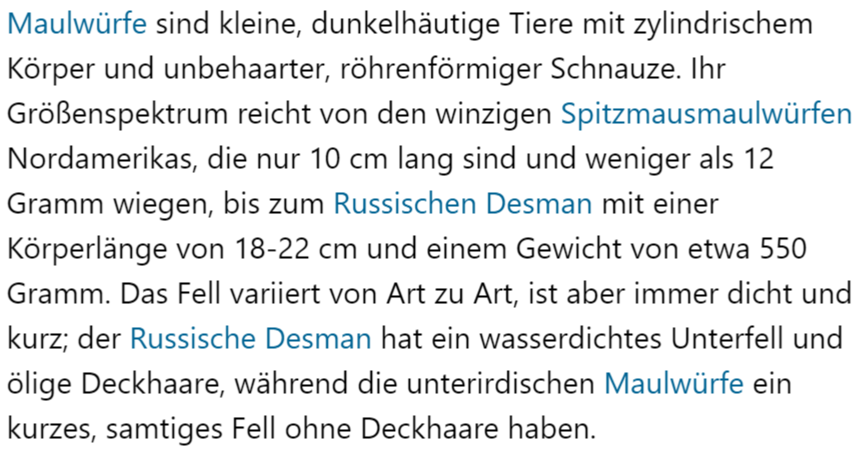
DE Target text

English - German

desman	Russischer Desman
mole	Maulwurf
Russian desman	Russischer Desman
shrew mole	Spitzmausmaulwurf

5) Conclusion and outlook

- The glossary feature seems to reduce post-editing effort
  - if the integrated target terms are correctly inflected
  - DeepL seems to be pretty reliable when integrating and inflecting nouns and adjectives
  - less reliable with verbs (Keller 2021:35)



Maulwürfe sind kleine, dunkelhäutige Tiere mit zylindrischem Körper und unbehaarter, röhrenförmiger Schnauze. Ihr Größenspektrum reicht von den winzigen Spitzmausmaulwürfen Nordamerikas, die nur 10 cm lang sind und weniger als 12 Gramm wiegen, bis zum Russischen Desman mit einer Körperlänge von 18-22 cm und einem Gewicht von etwa 550 Gramm. Das Fell variiert von Art zu Art, ist aber immer dicht und kurz; der Russische Desman hat ein wasserdichtes Unterfell und ölige Deckhaare, während die unterirdischen Maulwürfe ein kurzes, samtiges Fell ohne Deckhaare haben.

- Integration of the glossary feature in CAT tools
  - DeepL API supports the feature (e.g. see memoqdocs)
  - e.g., DeepL plug-in for Trados Studio not (yet)
  - Phrase Translate supports a glossary feature directly in the CAT tool, but the terms are not correctly inflected yet (Keller 2022:35)

depends on the engine/provider (e.g. DeepL and Phrase NextMT now often inflect terms correctly (April 2023))

Thanks for watching!

(references in the companion PDF)

References

DatCatInfo (n. d.): Welcome to DatCatInfo. *DatCatInfo*. https://datcatinfo.net/ (20 February 2023).

DeepL (2020): Customize DeepL Translator with the glossary feature. *DeepL blog*. https://www.deepl.com/en/blog/20200506 (20 February 2023).

DeepL (2021): Announcing glossary support for the DeepL API. *DeepL blog*. https://www.deepl.com/en/blog/announcing-glossary-support-for-deepl-api (20 February 2023).

DeepL (n. d.): Manage glossaries. *DeepL*. https://www.deepl.com/docs-api/glossaries/ (20 February 2023).

Intento (2021): The state of machine translation 2021. *intento*. https://try.inten.to/machine-translation-report-2021/ (20 February 2023).

Intento (2022): The state of machine translation 2022. *intento*. https://inten.to/machine-translation-report-2022/ (20 February 2023).

ISO 10897 (2019): *Terminology work and terminology science — Vocabulary*: Geneva: ISO Copyright Office.

References

ISO 12616-1 (2021): *Terminology work in support of multilingual communication — Part 1: Fundamentals of translation-oriented terminography*: Geneva: ISO Copyright Office.

Keller, Nicole (2021): DeepL integriert Glossarfunktion. In: edition 17(1), 34f.

Keller, Nicole (2022: Glossarfunktion in Phrase Translate. In: edition 18(2), 33–35.

Seidel, Jenny/Grützmacher, Ulrike (2020): Without terminology, translation is nothing. In: Dalla-Zuanna, Jean-Marc/Kurz, Christopher (Eds.): Translation quality in the age of digital transformation. Berlin: BDÜ Fachverlag, 184–199.

Wikipedia (31 January 2023): Talpidae. Wikipedia. https://en.wikipedia.org/wiki/Talpidae (20 February 2023).

Winter, Tom (2021): Terminologische Beeinflussung der Neuronalen Maschinellen Übersetzung: Ein Praxisbericht. In: edition 17(2), 5–10.

Winter, Tom/Zielinski, Daniel (2020): Terminologie in der neuronalen maschinellen Übersetzung. In: Porsiel, Jörg (Ed.): Maschinelle Übersetzung für Übersetzungsprofis. Berlin: BDÜ Fachverlag, 210–233.