Possibilities of Improved Terminology Adaption Features in Machine Translation Post-Editing

Case study from a German higher education institution

Abstract

As a freelance translator into English for several German higher education institutions (HEI), I face the challenge of aligning terminology with client decision-makers and authors. Furthermore, the widespread use of machine translation (MT) by staff at higher education institutions and the integration of terminology adaption into several MT engines has strengthened the business case for finding a common data model and data source for multilingual glossaries, which coincides with the aims of ISO/TC37 – Language and terminology.

Existing terminographic resources are currently not stored and shared consistently with all user groups, such as authors, translators and content editors, requiring them to rely on repetitive lexicographic research. While there are some organizational reasons, the technology is mature enough but not yet sufficiently integrated.

For example, computer-assisted translation (CAT) tools still struggle to adapt terminology automatically, even though several connected MT systems already support terminology adaption. This submission aims to illustrate the value of MT adaption to the machine translation post-editing (MT-PE)

workflow from a user perspective, hoping to convince stakeholders to prioritize terminographic features and develop some user recommendations for standards covered by ISO/TC37/SC 3.

Based on the argument that term validation is a significant effort, both in human translation (HT) and in machine translation post-editing (MT-PE), and that lack of a "single source of truth" and interoperability between computer assisted (CAT) tools and other terminology-consuming software (such as authoring tools) are the limiting factors, this submission provides an overview of terminology features in selected CAT tools, describes the development of a custom terminology management system, attempts to measure productivity gains from term adaption in different translation tools and proposes the term rate as the starting point for an additional indicator for productivity-relevant MT quality.

Contents

1. Introduction

a) Adoption of CAT in HEI translation

Harmonization of terminology across European HEIs has been described by (Ferraresi, 2017) as a barrier to international student mobility and communication between HEIs.[1] While claims that cheap, consistent translations into English will directly increase attraction and satisfaction of international students should be eyed with caution, the potential of terminology to simplify existing translation workflows should be examined.

To this end, I compare the existing terminology management features in some computer assisted translation (CAT) and machine translation. I outline the financial potential of terminology management compared to the current model of match discounts[2] by testing different methods of terminology adaption and relating the results to existing studies.

The term rate emerges as an additional quality indicator for MT output, which also provides information about the expected post-editing effort. However,

[1] Because the practice in Germany is that each HEI maintains its own terminology, creating a shared term base for all German HEIs is still out of scope. Ferraresi (2017) even finds that terminological variety "is especially the case in German-speaking countries, and in Germany in particular: here, one notices not only a lack of consensus across different universities, but also a lack of marked preferences within single universities."

[2] For an introduction into the match discount model, see (Dudi, 2016) or (Carl & Braun, 2018).

the realized speed gains will depend on variable factors such as term density and quality of the used term base.

Since existing productivity features in computer-assisted translation (CAT) are underutilized in the higher education sector, exploring other options seems worthwhile: According to a recent market survey (Ghamsharick, 2021), only 64% of 25 responding HEIs in Germany even use their own CAT tool, meaning that they are able to calculate match discounts without relying on an external provider.

Even so, 89.5% apply no match discounts at all. So while some HEIs even have dedicated staff for translation management, the core promise of CAT – the reuse of previous translations – is not leveraged, and neither is the promised savings potential.

The underutilization of CAT may be due to internal translation departments at HEIs lacking technical expertise or project management capacities. Another potential reason is that there are few repetitive segments, except in regulatory documents (see Table 1). So there might not be much leverage in recording and reusing segments, the main lexicographic feature of CAT tools.

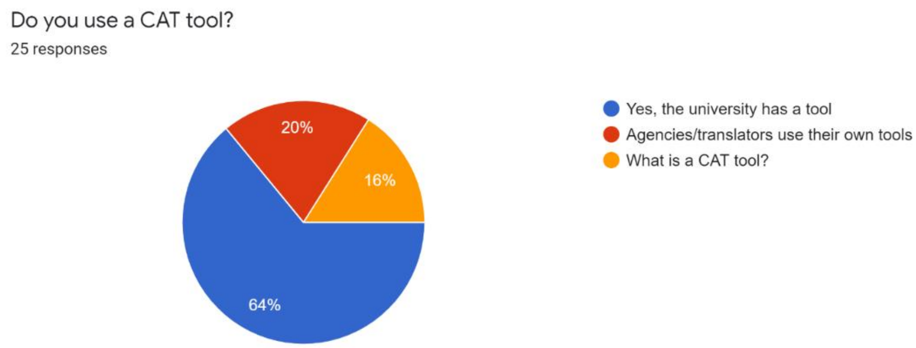


Figure 1 - Percentage of surveyed HEIs that (do not) use CAT tools. Source: (Ghamsharick, 2021)

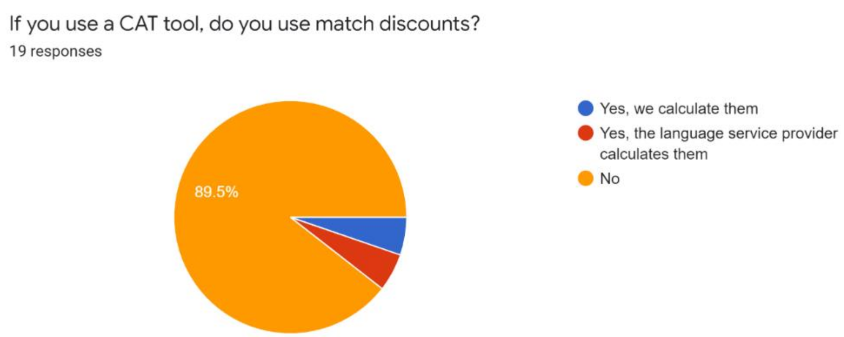


Figure 2 - Percentage of HEIs that (do not) apply match discounts. Source: (Ghamsharick, 2021)

While a well-maintained translation memory (TM) can be used to train custom machine translation models or to extract terminology, it is likely that the TM will serve mostly as a source for concordance searches to check how a term has been translated in the past, but this process is prone to propagating errors, producing inconsistent translations and still requires manual research effort. Actual productivity gains from using a TM have been studied by (Yamada, 2011), with inconclusive evidence.

With this in mind, it is understandable why 36% of the surveyed HEI end users in Figure 1 do not even have their own CAT tool. While many HEI staff members may be able translate into English, which is the only target language at most German HEIs, they may not be willing, able or authorized to use a CAT tool, especially when they do not see the value.

Hence, internal translators may resort to MT tools, like DeepL, especially for short texts. If translations are outsourced, the language service provider (LSP) / freelance translator may or may not use their own CAT tools, but even so, it is doubtful whether the expected savings really materialize.

This begs the question why CAT tools are still structured around the "segment recycling" use case, even though the value of term bases over TMs has been observed earlier. For example, (Garcia, 2014) writes:

> "Despite the emphasis traditionally placed on TMs, experienced users will often contend that it is the terminology feature which affords the greatest assistance. This is understandable if we consider that translation memories work best in cases of incremental changes to repetitive texts, a clearly limited scenario. By contrast, recurrent terminology can appear in any number of situations where consistency is paramount."

Yet, some technological changes are required to realize the full potential of terminology. A stronger business case for terminology adaption might help convince CAT and MT developers to prioritize terminology features, but reliable indicators are lacking.

b) Literature on terminology and translator productivity

The research on and measures for the quality of machine translation post-editing is quite rich, especially on differences between novices and professionals (such as (Daems, Vandepitte, Hartsuiker, & Macken, 2017)). Research also exists about the difference in terminological density between specialized and non-specialized texts, such as (Ferraresi, 2019).[3]

There is some research on the time translators spend validating target terms using sources such as search engines, bilingual websites, Wikipedia (lexicographic sources) compared to dictionaries and glossaries (terminographic sources). Interestingly, (Alvarez Lozano & Umana Corrales, 2020) in a study with screen recordings of professional and trainee translators, found that both groups have a clear preference for lexicographic sources and do not spend much time documenting their terminological choices. In other words, they prefer to validate terms manually instead of relying on dictionaries.

[3] It can be argued that a "specialized" text – in terms of term density – does not necessarily take longer to translate than a "non-specialized" text with a low term density. However, given the same term density, terminology management can impact MT-PE speed, since it reduces term validation effort, as discussed below.)

This would support the preference for TMs as a source for terminology over term bases, but this does not mean term bases should be neglected, but rather that they often *are* neglected in practice, rendering them unreliable.

Yet, it should be considered that a validated term base gives translation buyers more bargaining power, not just when setting prices, but also when defining quality standards. Terminological accuracy remains, arguably, a very visible criterion for translation quality, but without a reliable data source, there is no benchmark to measure against and hence no way to define accuracy.

Customers can have difficulty enforcing undocumented term preferences when there are multiple valid target terms in lexicographic resources, such as TMs or web searches. "Accurate but wrong" target terms are a problem when no term base is used, especially higher education translation, as shown in section 6.

c) Attempting to quantify match discounts

Since the effort spent on term validation is difficult to quantify, I attempted to first estimate the potential savings from "traditional" match discounts in my use case. For this, I checked the repeated translation units in my existing TM for the HEI.

The example in Table 1 *without* examination regulations compares potential savings from repetitions when excluding this high-match text type, while the lower example includes them.

Potential savings from match discounts without examination regulations			
	Word count	**No discount**	**-90% on words in repeated segments**
Overall wordcount in 2020/21	19.258	1.925,80 €	1.925,80 €
Of which in repeated segments	2.754	275,40 €	27,54 €
	Total	2.201,20 €	1.953,34 €
		Price difference	247,86 €
		Difference in %	11%
Potential savings from match discounts with examination regulations			
	Word count	**No discount**	**-90% on words in repeated segments**
Overall wordcount in 2020/21	51.369	5.136,90 €	5.136,90 €
Of which in repeated segments	12.491	1.249,10 €	124,91 €
	Total	6.386,00 €	5.261,81 €
		Price difference	1.124,19 €
		Difference in %	18%

Table 1 - Potential savings when applying match discounts to client's existing translation memory.

Source: own data

Using model figures, assuming that one word costs EUR 0.10 and a 90% discount is applied to all words in repeated segments, I calculated that price savings of 11% could be achieved on the existing translation memory when excluding high-word-count, repetitive examination regulation documents, and 18% when including this text type. (Even with different word prices, the differences in savings percentages should remain stable.)

While these discounts are not negligible, the fact remains that almost 90% of HEIs do not apply match discounts, as shown in Figure 2. This may also be owed to the complexity of the so-called Trados Discount Model (Dudi, 2016), and the difficulty of purchasing and deploying enterprise-grade CAT suites, which exceeds the capacities not just of HEIs, but also many small enterprises.

One argument against TMs by translators is that they allow the buyer side to reduce prices without increasing productivity (Yamada, 2011). Automation tools should focus on productivity first. If productivity really goes up, the unit price can be expected to fall eventually due to supply-side pressure.

Arguably the biggest boost to translator productivity has come from MT. This could be further increased by combining MT with a well maintained in-house glossary, but the existing software must better support this workflow.

Even more could be saved in cases of a single language combination with English as the target language by encouraging authors to translate their own texts using MT tools without using traditional CAT tools, thereby reducing the need for outsourced translations. This user group would also benefit from simpler translation tools and better terminology management, giving the client more choice of translation methods and more control over the outsourcing process.

2. Current state of terminology management

The TBX standard was introduced to solve the challenge of sharing documented terminological research between different tools (TBXinfo.net, 2021). My tests of built-in glossary features have shown that regardless of whether or not a TM is used, a non-customized NMT system, if combined with a well-maintained term base, can increase the time translators spend editing the output instead of validating terminology, but the lack of a common data exchange format is still a major barrier to terminological domain adaption in NMT.

Here again, choosing the right indicators matters. The focus on quality over productivity in evaluations of MT-PE remains a cognitive obstacle. The BLEU score is widely used to evaluate MT output (see, for example (Hu, Xia, Neubig, & Carbonell, 2021) or (Freitag & Al-Onaizan, 2016)). However, the unofficial "currency" in the language services industry, and the argument that buyers look for, is output over time – i.e. productivity – and not quality. As Daems et al. (2017) have noted:

> "Whereas the values given by such metrics [BLEU or METEOR] can be used to benchmark and improve MT systems, they [...] do not necessarily provide post-editors with valid information about *the effort that would be involved in post-editing the output*."

Alternative MT-PE quality metrics, like edit distance, still focus on delta between source and post-edited target text. As (Iizuka, 2019) writes:

> "The proposal to pay for post-editing by edit distance is based on a misconception: the idea that post-editing (and translation in general) is nothing more than typing. The key skill is to know what to type, and as a text becomes more specialized, *the post-editor's time is increasingly spent on validating meaning and less on actual keystrokes*."

Current CAT tools are designed to record keystrokes in translation memories and compare new source texts against these segment databases, while terminology is just an "extra feature" designed for ad-hoc management.

But terminology is also neglected on the business side. Language services industry experts, such as Beninatto and Johnson, in their "General Theory of the Translation Company" (2018) do not even classify terminology management as a support activity of a language service provider (LSP).

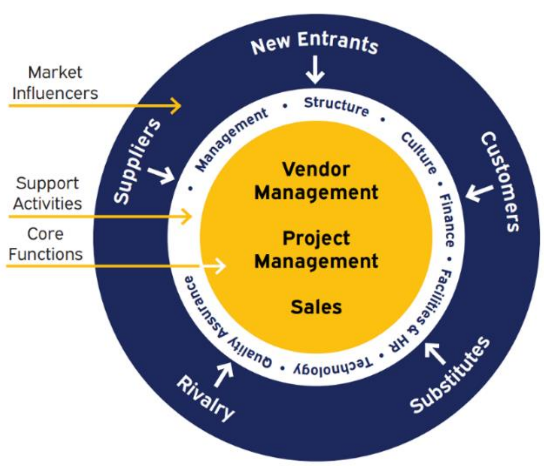


Figure 3 - Core Functions and support activities according to (Beninatto & Johnson, 2018)

Terminology management is usually left somewhere between quality assurance and technology. In practice, quality assurance means that another person (proofreader, project manager or quality manager) manually checks if terms defined in a glossary were used *after* the translation is done, because CAT tools are not able to insert the right terms at runtime. Validation of terms entered into the glossary is also frequently neglected, which means that the glossary may offer the wrong suggestions.

3. Who is responsible for terminology management?

While Kageura and Marshman (2020) describe a workflow for terminology management that begins with (automatic) terminology extraction, they do not mention which stakeholder begins this process. This indicates that either language industry experts have differing opinions on the value of and responsibility for terminology management or that they do not see terminology management as the responsibility of language service providers at all.

Terminologists such as Childress (2020) argue that terms should not be entered in a term base before the definition has been written by the author. Translators should only be responsible for finding target-language equivalents, but not for defining concepts.

Terminology is one of the most visible error categories in translation, and propagation of terminology errors can best be avoided by defining and discussing terms early on. Involving translators in the discussion or having them initiate talks is usually a challenge, if they are part of a separate team or work entirely outside of the organization. Leaving terminology management to translators effectively means making it an ad-hoc, downstream activity.

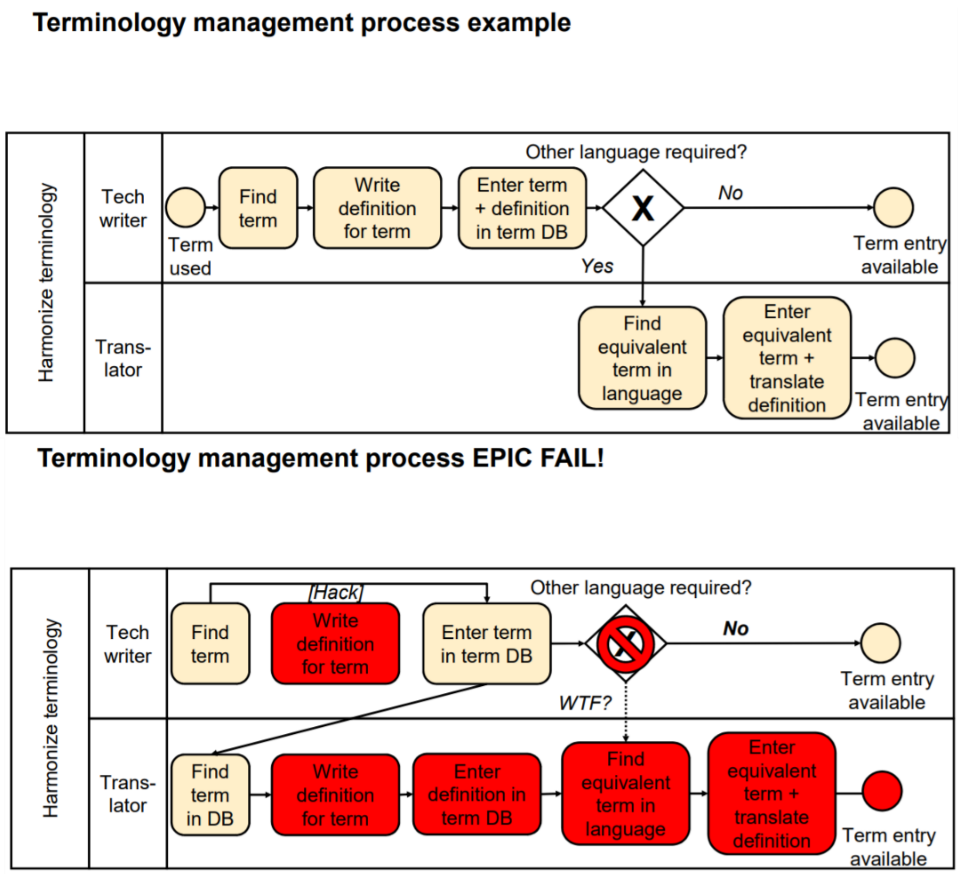


Figure 4 - Comparison of terminology management processes according to (Childress, 2020)

Facebook's Head of Terminology Uwe Muegge (2019) also advocates for clearly separating the process of terminology management from translation to reduce transition times spent on ad-hoc terminology management, but does not quantify the productivity gain. However, he correctly points out that term validation disrupts the translation workflow.

Hence, a term base can be expected to speed up the translation workflow, regardless of which tool is used. This conclusion is confirmed in section 6, provided that these tools can properly leverage the terminology, and here lies the technical challenge.

4. Finding a terminology management system for a HEI

The difficulty I faced in the HEI use case was finding a shared terminology management system (TMS) for authors, terminologists and translators, not necessarily the variety of data recorded. Even a tool that records terminological data with a great level of granularity is limited, if users cannot access and edit the terminology in their respective environments.

This corresponds to the description of dedicated TMS being difficult to integrate, as described by TerminOrgs in its Starter Guide (2016)**:**

TMS	Advantages	Disadvantages
Existing software (Excel, Google Sheets)	Easily available and to use collaboratively	Limited features, no integration or user management
Built-in features in CAT / authoring tools	Integrated into existing tools	Often on-premise installation, costs, limited terminology management, no integration with other tools
Dedicated TMS	Good terminology management	Difficult to integrate with translation/authorin g tools, costs

Table 2 - Types of TMS, adapted from (TerminOrgs, 2016)

Furthermore, the Starter Guide mentions some key features of a TMS, such as:

- Single repository: all terms must be in a single database
- Concept orientation: all info for a concept must be in in one entry
- Data elementarity: only one type of information per field
- Workflows: status concept to indicate the processing status of a term
- Quality assurance: ensuring that each term is reviewed and approved
- User management: making sure the reviewers have the required authorization
- Reporting: statistics and change overview
- System integration: ideally *automated* data exchange with authoring and translation tools

After researching some alternatives, a Google Sheet with advanced features added through Google Apps Script proved the most realistic option to meet the data elementarity requirements while enabling collaboration with the client organization. The red numbers in Figure 3 indicate how different spreadsheet features are used to meet the TerminOrgs requirements:

1. A simple status concept (New, Changed, Review)
2. Change tracking with script-enhanced cells that automatically capture the date when an entry is changed
3. Columns with context data categories, such as definition, comments, etc.
4. Column filtering to check all fields in a particular status for quick approval
5. XLSX or CSV export to import the latest version into a CAT tool

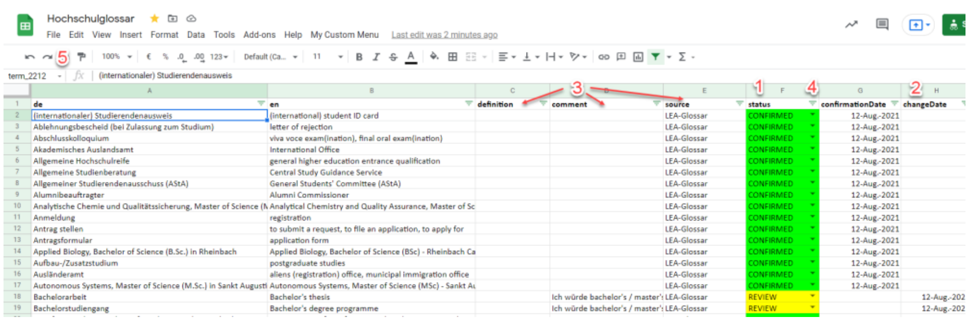


Figure 3 - Term base as a script-enhanced Google Sheets table. Source: own data

Yet, synchronization effort remains high, because the terminology cannot be automatically shared with CAT or authoring tools. However, this solution is more usable than some cloud-based, free terminology management tools that were tested (Terminologue.org, Lexonomy.eu), since they either use other file formats, such as customizable XML schemas (which theoretically would also accommodate TBX, but are difficult to configure for end users) or SQLITE.

Both XML and SQLITE are not easily editable in a spreadsheet tool or text editor. Using such "all-purpose editors" is still necessary to exchange data between the term database and various CAT tools, because despite initiatives to establish TBX as a common file exchange format for term bases, CSV and XLSX remain the unofficial standards, at least for cloud-native CAT tools, and there are few freeware TBX editors.

5. Current state of terminology exchange in CAT

As shown in Table 4, the more basic, cloud-native[4] CAT tools (Smartcat, Transifex, MateCAT) support either XLSX or CSV. Only Memsource and MemoQ, i.e., the more feature-rich, cloud-enabled software suites support TBX, while Trados uses a proprietary alternative.

The translation memory eXchange (TMX) format is also sometimes used as a workaround to exchange glossary data. TMX is also an easy way of training custom MT, such as ModernMT, since currently DeepL Translator seems to be the only NMT provider with terminology adaption and runtime (DeepL, 2020).

[4] Cloud-native refers to applications built to be used online, while cloud-enabled refers to software originally built to be installed on an operating system. Cloud-enabled CAT tools, such as Trados and MemoQ, tend to be more feature-rich, while cloud-native tools are easier to deploy due to subscription-based pricing (software as a service, SaaS) and no need for installation. See also (Rinner, 2016).

Table 4 shows that the prediction stated in (Rirdance & Vasiljevs) in 2006 has still not materialized, and the tendency in the newer generation of cloud-native translation tools seems to be fewer features and simpler formats:

> "It can be assumed that many developers of terminology management tools and other language processing applications will support TBX as an exchange format in the near future. Therefore TBX must be the recommended exchange format for terminological data in almost every specific interchange scenario."

	Standalone tool?	Remote/API access	Supports TBX	Import format	Export format	Multiple languag e pairs	Supports attributes	Morphology support	Autom. term adaption
CAT tools									
SmartCat	No	No	No	XLSX, MultiTerm XML	XLSX	Yes	Yes	No	No
MemSource	No	No	Yes	XLSX, TBX	XLSX, TBX	Yes	Yes	Word stem marker, fuzzy matching	No
Transifex	No	No	No	CSV	CSV	Yes	Yes	No	No
MateCat	No	No	No	TMX, XLSX	TMX, XLSX	Yes	No	No	No
MemoQ	Qterm	Yes	Yes	CSV, TBX, XLSX, MultiTerm XML	TBX, CSV, XLSX, MultiTerm XML	Yes	Yes	Fuzzy matching	During training (External adaptive MT)
Trados	MultiTerm	Yes	Proprietary alternative	MultiTerm XML	Multiterm XML	Yes	Yes	Fuzzy matching	During training (Own adaptive MT)
(Custom) MT tools									
DeepL Pro	Yes	No	No	CSV	None	No	No	No	Yes
ModernMT	No	No	No	TMX (workaround)	No	No	No	No	No

Table 4 - Overview of terminology management features in different CAT and MT tools. Source: own research and (Nimdzi, 2021)

It should be tested in practice what degree of data granularity is required for which terminological application. For example, a term base used in a CAT tool might not require part-of-speech information, while a term base used as a dictionary does. DeepL's own glossary feature does not support any metadata, although CSV supports between 255 and 1024 columns. Hence, when it comes to adding data categories, CSV is no less flexible than TBX.

In conclusion, despite the various proprietary terminology management tools that are available, for this use case, a custom solution based on XLSX remains the tool of choice. The HEI in this case has an annual translation budget between EUR 10,000 and 15,000, about 360 terms currently in the term bank, and only one language combination. The CAT tool I use is Smartcat, but the tool can be changed easily, since I only need to export my TMX and glossary as a CSV or XLSX file. The challenge of finding a dedicated terminology management system remains. For now, a spreadsheet remains the most usable option.

6. Quantifying productivity gains from term adaption

a) Testing for speed and accuracy

To test whether terminology improves speed across different tools, I wrote four nonsensical, but syntactically correct German sentences with a very high term density of approx. 25% (12 glossary terms out of 47 words):

> *Das **Institut für soziale Innovationen (ISI)** ist dem **Kanzler** und dem **Kuratorium** unterstellt.*
>
> *Das **Präsidium** prüft bei **Rückmeldung** eine **Studienarbeit**.*

> ***Vorübergehend Beschäftigte*** *beim* ***Vizepräsidenten für Innovation und Regionale Entwicklung*** *sind kein Teil der* ***Studierendenschaft****.*
>
> *Der Alumnibeauftragte weist* ***Lehrkräfte für besondere Aufgaben*** *in die* ***Verantwortlichkeiten*** *für ihre* ***Schwerpunktfächer*** *ein.*

I post-edited these sentences two times in the following translation tools:

- Smartcat, a cloud-native CAT tool which allows new projects to be set up comparatively fast and uses Google MT
- The neural MT tool DeepL Translator Pro
- The custom MT tool ModernMT

I post-edited them in each tool once without a glossary and once with the glossary activated to test for speed and accuracy gains, but also to test differences between translation tools (CAT and MT). The figures in Table 6 show the post-edited output with relevant terms highlighted in different colors.[5]

[5] In the case of ModernMT, the post-edited text appears in a separate MS Word instance, since the target text cannot be edited in the tool.

Productivity gain from terminology adaption in different translation tools						
	SmartCAT w/o term base	**SmartCAT with term base**	**DeepL w/o glossary**	**DeepL with glossary**	**ModernMT with TM w/o glossary**	**ModernMT with glossary**
Set-up time	00:00:40	00:00:43	00:00:10	00:00:17	00:00:30	00:00:24
Editing time	00:06:00	00:02:32	00:03:20	00:01:20	00:03:00	00:01:30
Total time	00:06:40	00:03:15	00:03:30	00:01:37	00:03:30	00:01:54
Verifiable correctly recognized terms	5	7	5	11	4	6
Verifiable incorrectly recognized terms	2	5	3	1	2	6
Unverifiable terms	5	0	4	0	6	0
Total terms	12	12	12	12	12	12

Table 5 - Productivity gains from terminology adaption in different tools. Source: own data

The green frames in the screenshots indicate that a term was correctly translated and did not need to be changed (**verifiable correctly recognized term** in Table 5). "Correct" with glossary activated means that the MT translation corresponds to the glossary entry. "Correct" without glossary activated means that I was able to manually verify that the target term was used on the HEI's website, i.e. using a lexicographic Google search.

A terminographic search in any non-organization-specific dictionary would not help validate these terms, because they are either organization-specific proper nouns (such as *Chancellor* and *The President's Office*) or common nouns that would return several options, even in a domain-specific dictionary. This high context sensitivity could also explain why both professional and trainee translators in the experiment described by (Alvarez Lozano & Umana

Corrales, 2020) strongly preferred lexicographic sources, such as Google searches on the client organization's website.

The red frames indicate that, in absence of a glossary, I was unable to verify a term without further consulting the client about their preferred usage. In practice, this would indicate the need to create an ad-hoc glossary entry.

In this test case, these **unverifiable terms** only occur in the test cases without glossary, because all terms occurring in the test sentences are already recorded in the glossary. This would not be the case in a real-life scenario, where new terms often need to be added ad hoc, even with a large existing glossary.

The yellow frames indicate that a term was verifiable but translated wrong by the MT system, i.e. it required manual correction (**verifiable incorrectly recognized term**).[6] The CAT tool, Smartcat, which uses Google MT, has a high rate of yellow frames, even with glossary activated, because glossary hits are not automatically inserted into the raw Google MT output. However, terms in this category are not inaccurate per se, they just do not match the preferred translation. Hence, they should not negatively affect the MT's overall quality rating.

[6] The figures show the edited MT output, so any "accurate but wrong" terms were already corrected manually.

Manual adaption of accurate, but non-preferred terms is also the current workflow in most CAT tools, which requires translators to spot relevant terms and insert glossary matches manually, a high-touch and error-prone process.

DeepL, on the other hand, recognized all but one of the terms, if the glossary was activated. ModernMT, where the glossary was fed into the custom MT model as a TMX file, only recognized half of the terms correctly. These two different approaches of integrating terminology into machine translation are described as "adaption during training" (ModernMT) or "adaption at runtime" (DeepL) (Eisold, 2021) and further explored in section c).

The disadvantage of adaption during training that became apparent in this test is that training can only increase the likelihood of a specific term being used, while adaption at runtime runs a separate process on the output to enforce target terms and can even react in real time to user changes.

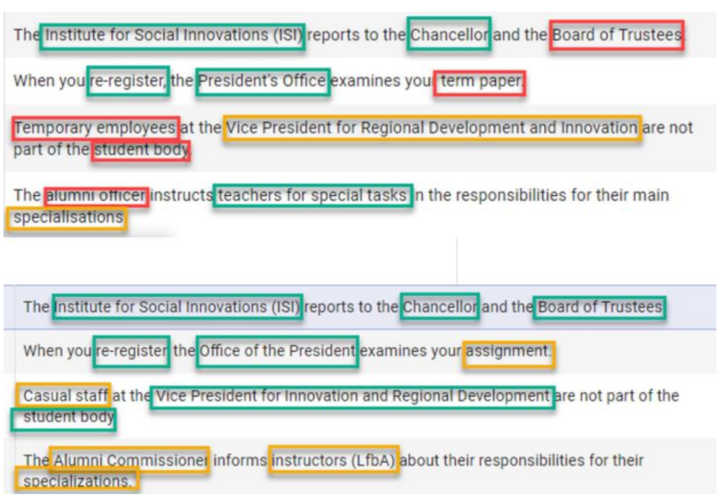


Figure 6 - Smartcat without and with glossary. Source: own data

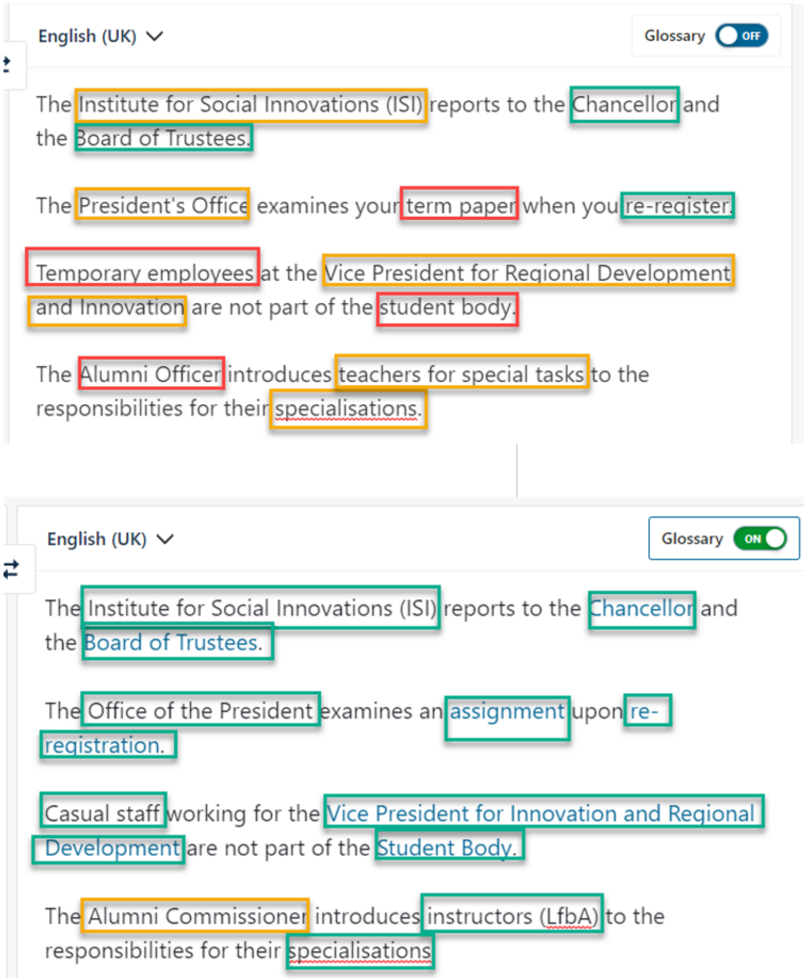
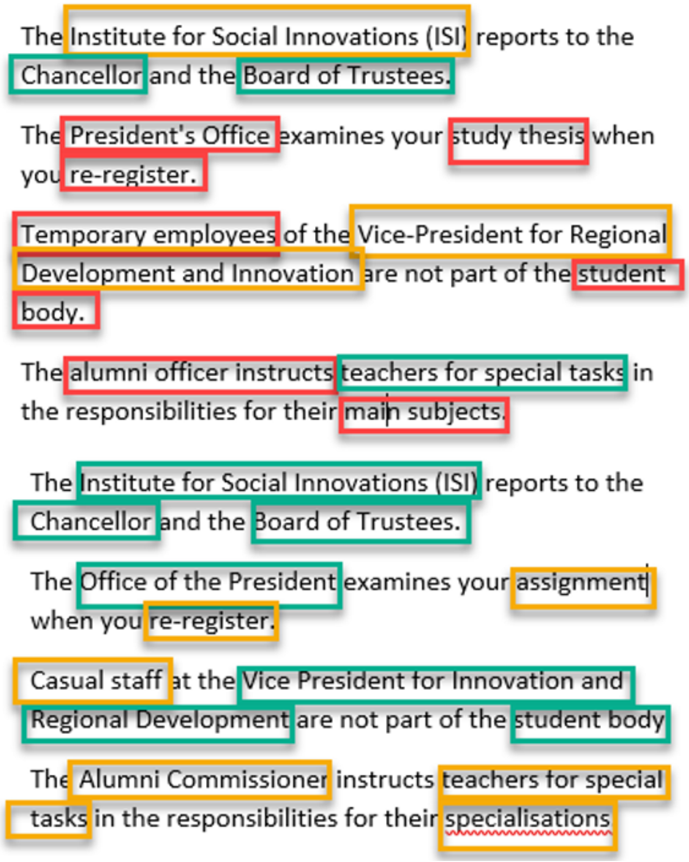


Figure 7 - DeepL without and with glossary. Source: own data



The Institute for Social Innovations (ISI) reports to the Chancellor and the Board of Trustees.

The President's Office examines your study thesis when you re-register.

Temporary employees of the Vice-President for Regional Development and Innovation are not part of the student body.

The alumni officer instructs teachers for special tasks in the responsibilities for their main subjects.

The Institute for Social Innovations (ISI) reports to the Chancellor and the Board of Trustees.

The Office of the President examines your assignment when you re-register.

Casual staff at the Vice President for Innovation and Regional Development are not part of the student body

The Alumni Commissioner instructs teachers for special tasks in the responsibilities for their specialisations

Figure 8 - ModernMT without and with glossary. Source: own data

b) Analysis

It emerged that terminology adaption at runtime, as used in DeepL, offers the greatest productivity gain in terms of speed and correctly recognized verifiable terms, also known as term rate (11 out of 12 = 91.6%). The alternatives, automatic terminology adaption during training (ModernMT), or manual terminology adaption during MT-PE (Smartcat / Google MT) also work faster if connected to a term base, but both lag behind in terms of correctly recognized verifiable terms (Smartcat: 7 out of 12 = 58.3%, ModernMT: 6 out of 12 = 50%).

The terms that are difficult for Google MT and ModernMT to recognize are common nouns with an organization-specific translation that differs from frequently used terms in the domain, such as *Schwerpunktfach* (preferred: specialisation), *Rückmeldung* (re-registration), or *Studienarbeit* (assignment). The results from ModernMT show that it is difficult to override this "noise" by using adaption during training.

The term *Lehkraft für besondere Aufgaben* demonstrates a difficulty in distinguishing between verifiable incorrectly recognized and unverifiable terms. In the cases without a glossary, the only verifiable source is the HEI's website, where this term is translated as "teacher for special tasks." Hence, if the output used this term, it was marked as correctly recognized, and if it used a different one, it was changed and marked as incorrectly recognized.

However, in the cases with glossary, the correct entry for this term was "instructor (LfbA)". Hence, what is "correct" depends on the validation source. This should be kept in mind when using the term rate as a quality measure for MT output – it can only be measured reliably, if the terminology is fed into the MT system.

It is also important to note that post-editing effort not related to term validation (i.e. rewriting) was minimal in this test case. Except for the second sentence, which suffered from terminological ambiguity (*Rückmeldung* can mean "feedback" or "re-registration / re-enrolment"), none of the MT systems had difficulties producing syntactically and semantically correct target sentences. Even the challenge of impersonal tone (*eine Studienarbeit*) was solved by all MT systems by inferring that the assignment belongs to the reader. Hence, the main difference between MT systems when it comes to perceived quality and MT-PE speed is the term rate.

Furthermore, a translation memory seemed to provide little added value in terms of domain adaption. The term base was the only means of adapting the NMT systems used in this study (CSV format for DeepL and TMX for ModernMT). If the German source text is sufficiently well-formed (S-V-O structure, no hidden agents, short sentences), most commercially available MT systems are capable of producing semantically accurate translations. Match discounts would not sufficiently capture these productivity gains.

Most of the manual effort went into validating terminology, which is why MT-PE was more than twice as fast in each of the systems, if some kind of terminology adaption (manual, at training or at runtime) was used. In other words, lack of a reliable term base increases the time a translator needs. How much exactly depends on the text's term density and other factors, such as file format, translation-conscious writing, CAT tool performance, etc.

Considering that Smartcat is one of the faster, cloud-native CAT tools, it is still considerably slower than the tested standalone MT tools (3:15 minutes vs. 1:37 for DeepL and 1:54 for ModernMT), even with glossary activated – for this particular test case.

c) Further testing of terminological accuracy in different MT tools

To compare the term rate between different terminology adoption methods in a larger text, I next counted only the number of correctly recognized terms (non-unique) in a larger sample of 525 words consisting of randomly selected segments from the existing TM.

This sample showed the term density statistics shown in Table 5. 47 terms (including of multi-word and non-unique terms) in a text of 525 words, equaling a term density of around 9% (or 6.48% unique terms), represents a use case closer to real life than the one in the previous section:

Term density in sample text of 525 words	
Total term density	47/525 = 8.95%
Unique terms as percentage of all terms	34/47 = 72.34%
Unique term density	34/525 = 6.48%

Table 5 - Term density in selected sample. Source: own data

I then machine translated this text in several engines and counted the correctly recognized terms (non-unique) in the raw MT output to calculate the term rate:

- DeepL with glossary
  - 39 / 47 terms = 82.98%
- ModernMT trained with TMX containing around 360 terms
  - 17 / 47 terms = 35.42%
- Google Translator,[7] untrained

[7] Since Smartcat uses Google MT, I translated the text sample directly in the Google Translate web interface.

  - 23 / 47 terms = 48.98%

- Google Auto ML[8] custom model trained with TMX containing a translation memory with over 8,000 segments and an additional TMX containing a glossary with over 360 terms
  - 23 / 47 terms = 48.98%

[8] Google Auto ML is a paid service that allows users to train custom MT models, similar to ModernMT.

SOURCE TEXT	TARGET TEXT	TERM RATE
Wenn der Studierende ein mündliches Kolloquium absolviert, sollte dies möglichst innerhalb von 2 Monaten, nachdem die Zulassungsvoraussetzungen erfüllt sind, stattfinden (Vgl. § 22 (1) MPO 2016).	If the student completes an oral colloquium, this should preferably take place within 2 months after the admission requirements have been fulfilled (cf. § 22 (1) MPO 2016).	3/3
Die Masterarbeit kann auch in Form einer Gruppenarbeit zugelassen werden, wenn der als Prüfungsleistung zu bewertende Beitrag des/der einzelnen Kandidat/en/in aufgrund der Angabe von Abschnitten, Seitenzahlen oder anderen Kriterien, die eine Abgrenzung ermöglichen, deutlich unterscheidbar und bewertbar ist und die Anforderungen nach § 19 (1) erfüllt.	The master's thesis can also be admitted in the form of a group project if the contribution of the individual candidate(s) to be graded as an examination result is clearly distinguishable and assessable due to the indication of sections, page numbers or other criteria that enable delimitation and fulfils the requirements according to § 19 (1).	5/5
Aufbauend auf den Kenntnissen wissenschaftlichen Arbeitens, welche im Rahmen der Lehrveranstaltung A1 (1. Semester) vermittelt wurden, werden die Studierenden nunmehr befähigt, zielgerichtet und unter Berücksichtigung verschiedener Quellen zu einem Thema zu recherchieren und dieses wissenschaftlich aufzubereiten.	Building on the knowledge of scientific work, which was taught in the course A1 (1st semester), the students are now enabled to research a topic in a targeted manner and to prepare it scientifically, taking into account various sources.	2/2
In der Rubrik "Optionen für das Auslandsstudium" können Sie sich einen Überblick darüber verschaffen, wie Sie ein Auslandsstudiensemester in Ihren Studienverlauf einbinden können.	In the section "Study abroad options" you can get an overview of how you can integrate a study abroad semester into your study schedule.	2/2
Für die Anerkennung als Schwerpunktfach suchen Sie alle fachlich zusammenhängende Kurse eines Fachgebietes aus dem Angebot der Gasthochschule heraus, die in dem Semester, in dem Sie vor Ort sind, angeboten werden (Beispiel International Management: International Trade, International Business Planning, International Economic Relation etc.).	For recognition as a specialisation, look for all subject-related courses in a subject area from the host institution's programme that are offered in the semester in which you are on location (example: International Management: International Trade, International Business Planning, International Economic Relation, etc.).	3/3
Teilnahme an mindestens zwei Dritteln der Gesamtdauer einer Lehrveranstaltung, sofern es sich dabei um eine Exkursion, einen Sprachkurs, ein Praktikum, eine praktische Übung oder eine vergleichbare Lehrveranstaltung (z.B. ein Planspiel) handelt.	Participation in at least two thirds of the total duration of a course, provided that it is an excursion, a language course, an internship, a practical course or a comparable course (e.g. a business simulation).	5/5
Wenn noch nicht alle Prüfungsleistungen erbracht wurden oder der Termin für das Kolloquium noch nicht feststeht, muss der Antrag zu einem späteren Zeitpunkt gestellt werden.	If not all examination results have been completed or the date for the colloquium has not yet been set, the request must be submitted at a later date.	3/3
Relevante Informationen (Kursinhalte, Kreditpunkte etc.) über die ausländische Hochschule an die Fachbereichsbeauftragten für das Auslandssemester weiterleiten; ggf. Learning Agreement erstellen.	Forward relevant information (course content, credit points, etc.) about the international H-BRS to the department coordinator for the study abroad semester; if necessary, prepare a Learning Agreement.	2/3
Diese richtet sich in erster Linie an die Studierenden des dritten Fachsemesters. Herzlich willkommen sind aber auch Studierende der anderen Semester.	This is primarily aimed at students in the third semester. However, students from other semesters are also welcome.	1/2
Zu Prüfenden dürfen nur die an der Hochschule Lehrenden und ferner in der beruflichen Praxis und Ausbildung erfahrene Personen, soweit dies zur Erreichung des Prüfungszwecks erforderlich oder sachgerecht ist, bestellt werden.	Only teachers at the H-BRS and persons experienced in professional practice and training may be appointed as examiners, insofar as this is necessary or appropriate to achieve the purpose of the examination.	1/1
In einer ersten Phase (2011 – 2014) umfasste die vorrangige Zielsetzung des Zentrums den strukturellen Aufbau des Instituts (z.B. Personalakquise, Entwicklung interner Governancestrukturen, Budgetplanung.)	In a first phase (2011 - 2014), the primary objective of the centre included the structural set-up of the institute (e.g. staff acquisition, development of internal governance structures, budget planning).	0/0
Während das Drittmittelvolumen erfolgreich gesteigert werden konnte, stellt die Publikationstätigkeit – auch aufgrund des hohen Anteils an mit Drittmitteleinwerbung gebundenen Personalressourcen - noch eine Schwachstelle dar.	While the volume of third-party funding was successfully increased, publication activity - also due to the high proportion of personnel resources tied to third-party funding - still represents a weak point.	0/2
Neben der Durchführung von nachhaltigkeits- und entwicklungsbezogenen Lehrveranstaltungen in den Fachbereichen der Hochschule, unterstützt das IZNE insbesondere auch die Initiative „Bildung für Nachhaltige Entwicklung“ durch die Beratung der Fachbereiche bei der Entwicklung entsprechender Studiengänge.	In addition to conducting sustainability and development-related courses in the departments of the H-BRS, the IZNE also supports the "Education for Sustainable Development" initiative in particular by advising the departments on the development of corresponding degree programmes.	2/4
Um den mit dem Wachstum des Zentrums einhergehenden gestiegenen administrativen Anforderungen sowie den sich aus der formulierten Zielsetzung ergebenden Anforderungen im Bereich Wissenschaftskommunikation begegnen zu können, ist eine weitere wissenschaftliche Mitarbeiterstelle notwendig.	In order to be able to meet the increased administrative demands associated with the growth of the Centre as well as the requirements in the field of science communication resulting from the formulated objectives, an additional academic staff position is necessary.	1/1
(2) Die Regelungen dieser Masterprüfungsordnung basieren auf dem Kooperationsvertrag zwischen der Deutschen Welle, der Hochschule Bonn-Rhein-Sieg und der Rheinischen Friedrichs-Wilhelm-Universität Bonn vom 13. Juni 2008, die den Masterstudiengang „International Media Studies“ gemeinsam verantworten und berücksichtigt ferner die Vereinbarungen der Hochschule Bonn-Rhein-Sieg mit weiteren Kooperationspartnern, die sich dem Studienprogramm anschließen.	(2) The provisions of these Master's Examination Regulations are based on the cooperation agreement of 13 June 2008 between Deutsche Welle, the H-BRS Bonn-Rhein-Sieg University of Applied Sciences and the Rheinische Friedrichs-Wilhelm-Universität Bonn, which are jointly responsible for the master's degree programme "International Media Studies", and also take into account the agreements of the H-BRS Bonn-Rhein-Sieg University of Applied Sciences with other cooperation partners who join the programme.	2/4
Die Urkunde wird von der Dekanin oder dem Dekan und von der bzw. dem Vorsitzenden des Prüfungsausschusses unterzeichnet und mit dem Siegel der Hochschule Bonn-Rhein-Sieg versehen.	The diploma is signed by the dean and the chairperson of the Examination Board and bears the seal of the H-BRS.	5/5
Diese Handreichung geht auf relevante, in der Durchführung der Prüfung abweichende Aspekte ein und beschreibt zudem den idealtypischen Ablauf einer Prüfung in diesem Setting - sowohl aus Sicht des Prüfenden als auch des Studierenden.	This handout addresses relevant aspects that differ in the conduct of the examination and also describes the ideal-typical course of an examination in this setting - both from the perspective of the examiner and the student.	2/2
	TOTAL TERM RATE	**39/47 = 82.98%**

Table 6 - Example: raw MT output and term rate from DeepL with glossary. Source: own data

The results confirm that DeepL with terminology adaption at runtime achieved the highest term rate (82.98%). ModernMT performed lowest (35.42%), while there was no difference between Google Translator and a trained Google machine translation system (48.98% for both).

Note that a high term rate does not measure overall MT quality, but my own comparison of the raw output in the different systems showed that little post-editing beyond manual term adaption was required in any of the tested systems.

With different MT systems being able to interpret a source text nearly equally well, the main distinguishing factor is these systems' ability to learn and adapt preferential term choices. This also has implications for data management, as there currently is no standardized way of feeding terminology into an MT system, just like there is no widely adopted standard for feeding terminology into CAT tools.

Regardless of the data source, when it comes to injecting terminology, different methods may produce similar results. Exel, Buschbeck, Brandt, & Doneva (2020), in their study on terminology constrained MT, which compares different methods of adding constraints, find that human translators saw no major difference in term accuracy between different terminology adaption methods (between 5.69 and 5.74 for *append-concat16* and *append-nofactors*, both for en-de and en-ru). As expected, terminology-constrained translations were rated, on average, around 25% higher for term accuracy than baseline for en-de and around 14% higher for en-ru.

More interesting even is that they find no big difference in overall translation quality ratings between adapted and non-adapted (baseline) translations. Translation quality was rated between 4.40-4.54 for en-de and 4.90-4.98 for en-ru, regardless of whether terminology was adapted.

This confirms my impression that terminology adaption does not affect text quality per se, but it does improve term accuracy benchmarked against a glossary (or term rate). As shown in Figures 7 to 9, untrained MT usually suggests accurate term translations, even if they are not the preferred choices. Unlike obvious mistranslations, "accurate but wrong" translations are difficult to eliminate without a benchmark.

	Term accuracy		Transl. quality	
	en–de	en–ru	en–de	en–ru
Baseline	4.52	4.99	4.40	4.90
Append-concat16	5.74	5.70	4.54	4.98
Append-nofactors	5.79	5.69	4.50	4.90

Table 7 - Results of human evaluation: term accuracy rating. Source: (Exel, Buschbeck, Brandt, & Doneva, 2020)

In this experiment, it appears that terminology adaption at runtime based on a manually validated term base offers the highest fidelity to a term base, but it remains to be seen whether terminology will be adapted directly in the MT system and then be sent to a CAT tool, or whether CAT tools will develop the ability to automatically adapt terminology at runtime.

7. Recommendations

The findings from these cursory experiments lead to the following recommendation for future standards in terminology management:

- ISO 16642:2017 - Computer applications in terminology — Terminological markup framework
    - Exchange of terminological data

In practice, XLSX and/or CSV are still the predominant formats for the exchange of term bases in translation tools, as seen in Table 4. One reason is also the lack of freely available converters and editors for TBX files. Terminology features within CAT tools are usually limited in their ability to batch edit term bases and to collaborate with authors, hence spreadsheet editors remain the tools of choice.

Any exchange format for term bases used by translators and authors must be easily readable and writable with commonly available tools. For small organizations, purchasing extra licenses for CAT, terminology management and authoring assistance tools and paying additionally for connectors, term base editors or term extraction tools is not just a bureaucratic effort in terms of market research, procurement and staff training, but also financially prohibitive.

When it comes to terminology adaption, the approach with the best accuracy seems to be adaption at runtime (see Table 5). This is possible with any CAT tool with a glossary feature and machine translation, however, it currently needs to be done manually. Another approach would be to connect a custom MT system and pre-train it with glossary data, but adaption at training is prone to interferences. In either case, automated terminology adaption should replace both manual adaption during translation and post-translation terminology QA checking as the methods of choice.

- ISO 26162-1:2019 – Management of terminology resources — Terminology databases — Part 1: Design
    - Terminology database design for distributed, multilingual terminology management

Maybe even more important than the format is the question where to store the term bases. Self-made solutions using spreadsheets are still prevalent, due to the ease of editing data with a spreadsheet editor and sharing data with authoring or translation tools.

While self-made solutions do not connect automatically to other tools, the same goes for dedicated TMS. Combined with the difficulty of implementing and learning new tools, this poses a challenge for the wider adoption of software that can handle terminology-specific file formats.

The current terminology workflow in CAT is shown in Figure 9. Even in advanced CAT suites, such as SDL Trados and MemoQ, which offer stand-alone terminology software, only the manual import between terminology management system and glossary component disappears, because the glossary component *is* the terminology management system. Any kind of terminology adaption, whether during training or at runtime (within the CAT tool) still requires manual configuration.

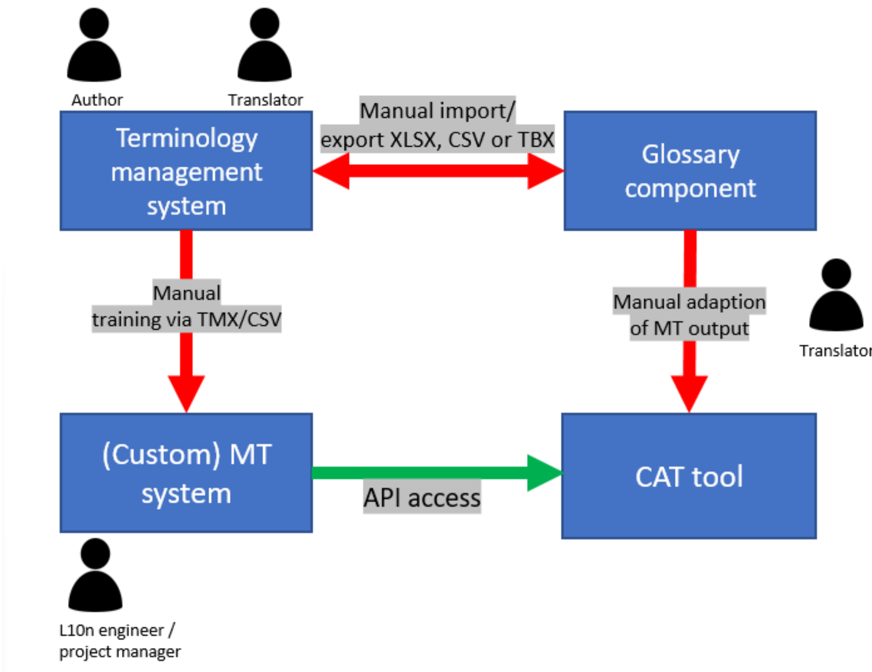


Figure 9 - Current terminology workflow in CAT tools. Source: own data

What should be kept in mind is that terminology must already be collected and defined during authoring (see Figure 4), especially when it comes to common nouns that have a preferred target term within the organization, such as, *request* (not *application*) as the preferred translation for *Antrag*). Existing translation memories as well as monolingual corpora can be used to extract terminology in advance.

Furthermore, many short texts are translated ad-hoc, by HEI staff members who may lack access to automated terminological resources. While ad-hoc translation is a good way to leverage organizational knowledge (e.g., when authors translate their own texts) and lower external translation costs without reducing output, terminology tools should be designed to enable translations outside of CAT tools as well.

- ISO 22128:2008 – Terminology products and services — Overview and guidance
  - Guidance for work contracts in the field of terminology

Speaking for the public higher education sector in Germany, even CAT tools are not used universally (see Figure 1), although many institutions translate their content into English. Software procurement is difficult due to lack of qualified experts on the buyer side and contractual design issues. To create the term base discussed in this submission, I negotiated a separate purchase order for terminology research, although my framework contract is for translation services.

Terminology work can quickly cross over into software development or consulting, where pricing and scope are notoriously difficult to quantify. It can take some effort just to find out that an intended outcome is not achievable.

For example, while exploring the available solutions for terminology management, both CAT-integrated and stand-alone tools, I found that while there are many feature-rich terminology tools available, they would not be able to automate the crucial steps in Figure 9. So instead of buying a tool that doesn't do exactly what I need, I used a free one that doesn't either.

Data access, authorizations and organizational silos can also be impediments for contractors working in the field of terminology. My main partners within the studied HEI were content managers, and terminology management is not their core focus. The available glossary had been written and "locked away" on the intranet several years ago and not been updated since. While there are some parts of terminology that can be outsourced, such as software consulting or data conversion tasks, terminology maintenance is everybody's – and therefore nobody's – core responsibility.

One interesting initiative to centralize linguistic resources for HEIs on the sub-national level is BaySev (BaySev, 2021), the "Bavarian Service Center for English-language Administrative Documents at Higher Education Institutions". However, these resources are only accessible to official HEI employees, and translators are often freelancers and therefore have no access to existing TMs and TBs.

In summary, terminology management should unfold increasing value as CAT tools migrate into the cloud and MT-PE proceeds to replace match discounts. The present attempt to quantify the value of terminology should help bring more attention to the importance of developing solutions on the technical side.

The term rate, here understood as the percentage of correctly recognized terms defined in a term base, as a more specific measure of term accuracy, should be further explored as an indicator for productivity in raw MT output, although the exact correlation remains to be quantified. A shift from quality and pricing towards productivity would also help better align interests between (translation) software developers and end users.

References

Alvarez Lozano, N., & Umana Corrales, O. (2020). Levels of involvement in terminology practice: differences and similarities in a group of trainee translators and a group of professional translators. *Revista Espacios*, 1 ff.

BaySev. (2021, August 24). Retrieved from Bayerische Servicestelle für englischsprachige Verwaltungsdokumente an Hochschulen: http://www.baysev.de

Beninatto, R., & Johnson, T. (2018). *The General Theory of the Translation Company.* Nimdzi.

Carl, M., & Braun, S. (2018). Translation, interpreting and new technologies. In K. (. Malmkjaer, *The Routledge Handbook of Translation Studies and Linguistics* (p. 375). New York, Oxon: Routledge.

Childress, M. (2020, February 25). Terminology management at SAP. Retrieved August 18, 2021, from https://www.technical-communication.org/fileadmin/user_upload/20200225_org_tekom-dk_term-mgmt-SAP_public-reduced.pdf

Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017, August 02). Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort. *Frontiers in Psychology, 8*. Retrieved 08 10, 2021, from https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01282/full

DeepL. (2020, May 6). *Customize DeepL Translator with the Glossary*. Retrieved from DeepL.com: https://www.deepl.com/en/blog/20200506

Dudi, P. (2016, March 27). *Word Counts, The Trados Discount Model & Weighted Words.* Retrieved from LinkedIN: https://www.linkedin.com/pulse/word-counts-trados-discount-model-weighted-words-petro-dudi/

Eisold, C. (2021). *https://exceltranslations.com/terminology-machine-translation/.* Retrieved from Excel Translations: https://exc

eltranslations.com/terminology-machine-translation/

Exel, M., Buschbeck, B., Brandt, L., & Doneva, S. (2020). Terminology-Constrained Neural Machine Translation at SAP. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 271–280). Lisboa: EAMT. Retrieved from https://aclanthology.org/2020.eamt-1.29/

Ferraresi, A. (2017). *Terminology in European university settings. The case of course unit descriptions.*

Ferraresi, A. (2019). HOW SPECIALIZED (OR POPULARIZED)? Terminological density as a clue to text specialization in the domain of food safety. *Lingue e Linguaggi*, 17-39.

Freitag, M., & Al-Onaizan, Y. (2016). Fast Domain Adaptation for Neural Machine.

Garcia, I. (2014). Computer-Aided Translation. In S.-W. Chan, *The Routledge Encyclopedia of Translation Technology.* Routledge.

Ghamsharick, E. (2021, 05 17). *Market survey for university translation in Germany*. Retrieved from Germling.com: https://germling.com/2021/05/17/market-survey-for-university-translation-in-germany/

Guy Champagne Inc. (2004, April 30). *The Economic Value of Terminology*. Retrieved from Danterm.dk: https://www.danterm.dk/docs/EconomicValueTerminology-1.pdf

Hu, J., Xia, M., Neubig, G., & Carbonell, J. (2021). Domain Adaptation of Neural Machine Translation by Lexicon Induction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2989-3001). Florence: Association for Computational Linguistics.

Iizuka, I. (2019, March 13). *Edit Distance: Not a Miracle Cure.* Retrieved from RWS.com: https://www.rws.com/blog/edit-distance-not-a-miracle-cure/

Kageura, K., & Marshman, E. (2020). Terminology extraction and management. In M. (. O'Hagan, *The Routledge Handbook of Translation and Technology* (p. 61 ff.). London: Routledge.

Muegge, U. (2019, November 26). *Why You Should Care about Terminology Management—Even if You Never Translate a Single Term.* Retrieved from The ATA Chronicle: https://www.ata-chronicle.online/featured/why-you-should-care-about-terminology-management-even-if-you-never-translate-a-single-term/#sthash.b83dmRcB.dpbs

Nimdzi. (2021, August 12). TMS Feature Comparison Tool. Retrieved from https://www.nimdzi.com/tms-compare?updated=true

Pawelec, M. (2018, 03 29). *The Case of Skipping QA.* Retrieved from MemoQ Blog: https://blog.memoq.com/the-case-of-skipping-qa

Rinner, I. (2016). Cloudbasierte versus desktopbasierte Translation Memory Systeme: SDL Trados Studio 2015 und MemSource im Vergleich. Vienna. Retrieved from https://termcoord.eu/wp-content/uploads/2013/08/Iris_Rinner_MA.pdf

Rirdance, S., & Vasiljevs, A. (. (2006). *Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project.* . Riga: Tilde.

TBXinfo.net. (2021, March 13). *Introduction to TermBase eXchange (TBX)*. Retrieved from TBXinfo.net: https://www.tbxinfo.net/tbx-about/

TerminOrgs. (2016). *Terminorgs.net.* Retrieved from TerminOrgs Starter Guide: http://www.terminorgs.net/downloads/TerminOrgs_StarterGuide_V2.pdf

Vezzani, F., & Di Nunzio, G. (2020). On the Formal Standardization of Terminology Resources: The Case Study of TriMED. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 4903–4910). Marseille: European Language Resources Association (ELRA).

Wilkinson, M. D. (2016, March). *The FAIR Guiding Principles for scientific data management and stewardship.* Retrieved from National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/26978244/

Yamada, M. (2011). The effect of translation memory databases on productivity. *Translation research projects 3* , 63-73.