Deutscher Terminologie-Tag e.V.

17. Jahrgang | www.dttev.org

edition

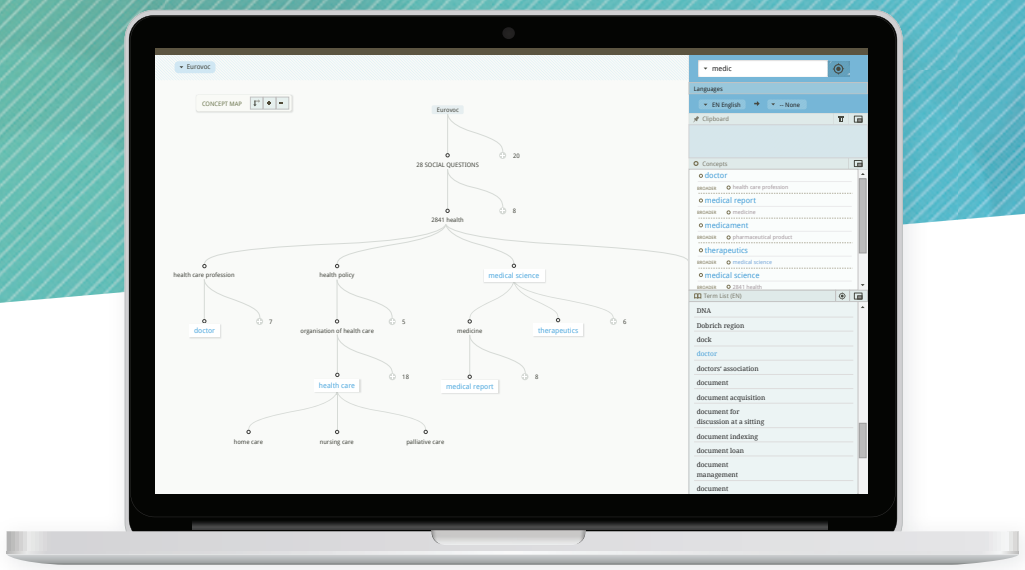
Fachzeitschrift für Terminologie

2|21

Terminologie nimmt weiter Fahrt auf mit NMT

Your Multilingual Knowledge System

Extreme Visual Terminology



360-degree control through systematic approach: **multilingual Concept Maps**

Proven enterprise deployments, scalable with Single sign-on: appealing, **plug-in free browser solution**

Enable NMT workflows, tune enterprise search, annotate and classify texts: **versatile integrations**

coreon

Knowledge meets language.

coreon.com

„Das Thema Gendern boomt!“

Dies ist ein Zitat von der tekom-Herbsttagung 2021, bei der diesmal ein ganzer Vortragsstrang zum Thema „Gendern“ auf dem Programm stand. Das Thema wurde dort für die Technische Redaktion von allen Seiten beleuchtet. Gerade in der gesprochenen Sprache klingt uns das Gendergerechte oft seltsam in den Ohren. Regeln, Normen, Vorschriften gibt es noch nicht, aber der Sprachgebrauch befindet sich bereits im Wandel. Die Organe, die Regeln aufgrund der Beobachtung des Sprachgebrauchs machen (DIN, deutscher Rechtschreibrat), haben noch keine Empfehlungen ausgegeben. Nur der Duden hat angefangen und neue weibliche Berufsbezeichnungen aufgenommen. Das Gendern wird daher für die Terminologiearbeit zunehmend relevant. Erste Gedanken dazu finden Sie in den Beiträgen „Das Gendern – terminologisch betrachtet ein Missverständnis mit Skandalpotenzial“ und „Ist Gott weiblich?“ in diesem Heft.

Auch das Thema Terminologie in der neuronalen maschinellen Übersetzung lässt uns nicht los. Ergänzend zu den Schwerpunktthemen der letzten Ausgabe berichtet Tom Winter hierzu aus der Praxis.

Die Evaluierung von Termbanken ist ebenfalls ein aktuelles Thema, dem sich Beate Früh und Dóra Mária Tamás in ihrem Beitrag „Quality Evaluation of Termbases” widmen.

Die Rubrik Tools & Ressourcen füllt Nicole Keller mit aktuellen Informationen zur computergestützten Terminologieprüfung.

Unter Wissenswertes zeigen Franziska Fischer und Pascal Müller in Ihrem Beitrag „Wasser oder Wässer – ein Fall für den Terminologen?“ auf, warum Sie die Datenkategorie „Numerus“ in der Terminologiedatenbank nicht stiefmütterlich behandeln sollten.

Ankündigungen von nationalen und internationalen Konferenzen runden das Heft ab.

Wir wünschen Ihnen wie immer eine vergnügliche und interessante Lektüre und freuen uns über Ihre Rückmeldungen.

Angelika Ottmann
Redaktionsleitung
redaktion@dttev.org

Dr. Annette Weilandt
Redaktionsleitung
redaktion@dttev.org

Inhalt

Editorial

Themen

Tools & Ressourcen

Wissenswertes

17. Jahrgang.
edition erscheint zweimal im Jahr (Juni/Dezember).
ISSN 1862-023X.
Herausgeber: Deutscher Terminologie-Tag e.V. (DTT) – www.dttev.org
Redaktionsleitung: Angelika Ottmann und Dr. Annette Weilandt – redaktion@dttev.org
Titelseite, DTT-Anzeigen, Layout und Satz: Tamara Arndt – layout@dttev.org
Anzeigen: Olga Buchstaller-Vodopiyanova – anzeigen@dttev.org
Lektorat: Christin Sonnberger
Mediadaten und komplettes Impressum: www.dttev.org/edition/impressum

Terminologische Beeinflussung der Neuronalen Maschinellen Übersetzung

Ein Praxisbericht

Tom Winter

There are various ways to influence the output of neural machine translation. Using two application examples from Deutsche Bahn AG, the article shows the methods of data integration, describes examples of preparation and processing of terminological data, and explains evaluation methods of the NMT output.

Keywords: neural machine translation, NMT, NMT training, NMT customization, terminology, morphosyntactic terminology integration, error classification, quality evaluation

Ein Lieblingswitz einer Kollegin lautet: „Wie viele Übersetzer braucht es, um eine Glühbirne zu wechseln? – Antwort: Es kommt auf den Kontext an." Und da die Digitalisierung zwar Einiges leichter macht, aber an der Komplexität von Sprache nicht viel ändern wird, führt sich dieser Witz auch in Bezug auf den Einsatz von Neuronaler Maschineller Übersetzung (NMT) fort: „Wie groß ist der Aufwand für die Beeinflussung von NMT-Output? – Antwort: Es kommt auf den Kontext an."

NMT-Training bzw. NMT-Customization (Anpassung) bietet großartige Möglichkeiten, den NMT-Output in Richtung der Anforderung zu bewegen, für die der Output einer generischen Engine nicht genügt. Gründe für Training und/oder Customization von Übersetzungsmodellen können mannigfaltig sein – ebenso mannigfaltig ist die Erwartung an den Output und ebenso mannigfaltig sind die verschiedenen verfügbaren Systeme und Möglichkeiten der Beeinflussung. Manchmal ist der Einsatz von Millionen konsistenter Satzsegmente notwendig und manchmal genügt ein simpler Glossar-Upload.

Die Deutsche Bahn AG verwendet NMT in diversen Kontexten. Ob die NMT beeinflusst wird und in welcher Form dies geschieht, ist stets abhängig von folgenden Faktoren:

- Anforderungen an Qualität (z. B. durch Sicherheitsrelevanz, Außenwirkung etc.)
- Komplexität des Projektrahmens (Stil, Idiomatik, Terminologie)
- Datenverfügbarkeit (projektspezifische bilinguale Daten und Terminologie)
- Budgetrahmen (der Kreislauf aus Datenaufbereitung, (Re-)Training und Evaluierung kann mehrere Zyklen umfassen und schnell zum Kostentreiber werden)
- zeitlicher Rahmen (der Kreislauf aus Datenaufbereitung, (Re-)Training und Evaluierung kann mehrere Zyklen umfassen und sehr lange dauern)
- technische Möglichkeiten (seitens des NMT-Anbieters)

Training vs. Customizing

In der Regel gibt es zwei Möglichkeiten, den Output von NMT zu beeinflussen:

- Training = das Erstellen eines neuen Übersetzungsmodells „from scratch" und
- Customization (auch: Domain Adaptation) = die Anpassung bestehender, generischer Übersetzungsmodelle durch kunden- oder projektspezifische Übersetzungsdaten

Auch beim Training gibt es Modelle, bei denen eine Priorisierung des Datenmaterials möglich ist. Dabei wird in der Regel zunächst ein generisches Datenmodell trainiert (z. B. unter Verwendung frei zugänglicher Korpora), auf welches im Anschluss spezielles, projektspezifisches Material aufgesetzt wird. Vorteil dieses Vorgehens: Durch das Eigentraining des generischen Modells kann bereits die Grundqualität beeinflusst werden. Nachteil: Das Trainieren

eines grundneuen Modells verlangt große Datenmengen (ca. 1 Million Segmentpaare).

Als Grundsatz für jedes erfolgreiche Training gilt: Die ausgangssprachliche Seite des bilingualen Trainingsmaterials (Korpus wie Glossar) muss variantenreich gestaltet sein, wohingegen die zielsprachliche Seite vor allem konsistent sein muss. Besondere Krux dieses Gesetzes: Trainingskorpus wie -glossar müssen immer für jede Sprachrichtung separat aufgebaut werden.

Direktes vs. indirektes Sprachpaar

Einige NMT-Provider arbeiten mit sogenannten indirekten Sprachpaaren. Dabei wird die Ausgangssprache zunächst in eine Pivotsprache übertragen und aus dieser dann in die Zielsprache übersetzt. Als Pivotsprache wird in den meisten Systemen Englisch verwendet.

Ein typischer Fehler, der auf die Verwendung einer Pivotsprache hindeutet, ist die Pronomenverschiebung durch „you“, das im Englischen sowohl die 2. Person Singular als auch die 2. Person Plural repräsentiert:

„Ich liebe Sie.“ -> „I love you.“ -> „Je t‘aime“ [Ich liebe Dich]

Diese Form ist risikobehaftet, da die Zwischenübersetzung einen unkontrollierbaren Eingriff in die Übersetzung darstellt, der sehr häufig negativen Einfluss auf die Übersetzung der Zielsprache hat. Für das Customizing empfehlen sich daher ausschließlich direkte Sprachpaare.

Korpusbasiert vs. glossarbasiert vs. korpus- und glossarbasiert

Die terminologische Beeinflussung von NMT kann korpusbasiert, glossarbasiert und in kombinierter Form stattfinden.

Die korpusbasierte Methode entspricht einer statischen Integration der Terminologie im Rahmen des Trainings bzw. des Customizings, da die Terminologie ausschließlich im Kontext des Trainingsmaterials integriert wird. Für eine erfolgreiche Schulung sind eine genügende Anzahl an Segmenten, eine genügende Anzahl der Terminologiepaare in unterschiedlichen Satzkonstellationen und absolute Konsistenz in der Verwendung ausschlaggebend. Die konkreten notwendigen Mengenangaben unterscheiden sich je nach Beeinflussungsform und Anbieter. Bei dieser Methode kann nicht gewährleistet werden, dass die Terminologieintegration in jedem Fall greift. Vorteilhaft ist allerdings die wahrscheinlichere Erkennung und Ersetzung durch fremdsprachliche Äquivalente auch bei flektierten Formen.

Die glossarbasierte Methode ist die dynamischere Form der Terminologieintegration, da für deren Einbindung nicht die gesamte Engine neu geschult werden muss. Das Glossar darf keine Homonyme enthalten und sollte entsprechend der Grundregeln aufgebaut sein: ausgangssprachlich variantenreich, zielsprachlich konsistent. Der Einsatz glossarbasierter Terminologieintegration steigert zwar die Wahrscheinlichkeit der Terminologieerkennung und -ersetzung durch Äquivalente in der Zielsprache, allerdings basiert der Erkennungsmechanismus auf reinem Zeichenkettenabgleich. Sofern im Glossar keine flektierten Formen aufgeführt sind, werden also nur Grundformen erkannt und übersetzt. Auch kann die stumpfe Integration der Grundform die grammatikalische Struktur des zielsprachlichen Satzes schädigen.

In einem Interview in der edition 01/2021 unterscheidet Samuel Läubli bei der technischen Umsetzung der Terminologieintegration zwischen Maskierung (relativ sichere Erkennung und Übersetzung der Terminologie in ihrer Grundform) und Provokation (grundsätzlich weniger sichere Erkennung der Terminologie, allerdings auch Erkennung und Ersetzung von flektierten Formen). Diese beiden Formen seien an dieser Stelle um zwei Formen erweitert:

Zum einen lässt sich die Funktion der Maskierung durch die manuelle Erweiterung des Glossars um flektierte Formen optimieren, sofern dies im Rahmen der Sprachkombination als möglich und sinnvoll bewertet wird. Der Vorgang und Aufwand mag im ersten Schritt an einen Rückfall in regelbasierte Zeiten erinnern, allerdings, nüchtern betrachtet, besteht unsere Rolle im Umgang mit der Künstlichen Intelligenz wohl immer aus dem Zweiklang aus Optimierung der Regel und (zwischenzeitlichem) Auffangen der Ausnahme. So betrachtet führt das Auffangen der Ausnahme, die erweiterte Maskierung, in diesem Fall zu einem optimierten Ergebnis.

Befehlsblock	carnet d'ordres
befehlsblock	carnet d'ordres
Befehlsblöcke	carnets d'ordres
befehlsblöcke	carnets d'ordres
befehlsblocks	carnet d'ordres
Befehlsblocks	carnet d'ordres

Abb. 1: Glossar mit flektierten Formen

Und den Glauben daran, dass dieser Aufwand tatsächlich nur temporären Charakters ist, befeuert die neueste Entwicklung von DeepL mit der Etablierung der morphosyntaktischen Terminologieintegration. Erkennung und Integration beziehen sich dabei nicht nur auf die Benennung an sich, sondern schlagen sich auf die komplette Zielsatzgrammatik nieder, indem vom Genus- oder Numeruswechsel betroffene Wortarten, z. B. Artikel und Adjektive, ebenfalls angepasst werden.

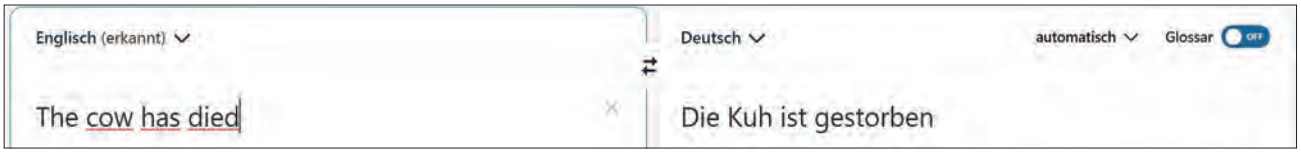


Abb. 2: Übersetzung ohne benutzerdefiniertes Glossar (Quelle: DeepL.com; Zugriff am 21.09.2021)

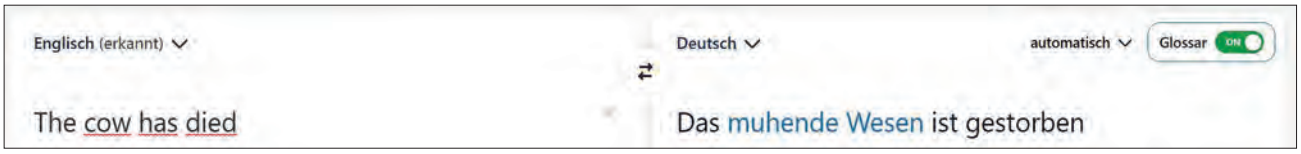


Abb. 3: Im Glossar ist „cow“ als „muhendes Wesen“ hinterlegt (Quelle: DeepL.com; Zugriff am 21.09.2021)

Glossaraufbereitung

Die Glossarerstellung bedeutet eine Verlagerung der hoheitlichen terminologischen Aufgaben aus der Begriffsbezogenheit heraus in die Niederungen der Benennungsorientiertheit. Umso wichtiger ist daher die korrekte Glossarerstellung, die neben der Auswahl der passenden Begriffe nur zwei Regeln folgen muss:

1. Ausschluss mehrdeutiger Benennungen in der Ausgangssprache zur Vermeidung der Erzwingung falscher Ersetzungen in der Zielsprache
2. Integration aller eindeutigen Benennungen in der Ausgangsprache (inkl. verbotener Benennungen) und konsistente Ersetzung durch einheitliche Zuordnung zum jeweilig bevorzugten Äquivalent in der Zielsprache

Abbildung 4 veranschaulicht die Umwandlung begriffsbezogener Einträge in benennungsbezogene Glossare unter Beachtung oben beschriebener Regeln.

Die Auswahl der Trainingsdaten ist stets ein Wandeln im Graubereich: In den meisten Fällen muss das Trainingsmaterial sehr präzise auf den spezifischen Use Case zugeschnitten sein. Es ist also in den seltensten Fällen sinnvoll, einfach alle vorhanden Translation Memorys (TM) oder die gesamte Terminologiedatenbank zu verwenden. Im Zentrum steht allerdings weniger die Fachsprachlichkeit als die Frage, welche Ersetzung in der Zielsprache in jedem Fall erzwungen werden soll bzw. muss. Dabei kann es auch sinnvoll sein, allgemeinsprachliches Vokabular oder Phrasen in das Glossar zu integrieren, um wiederkehrende Fehler zu vermeiden. Abbildung 5 zeigt ein Beispiel zur Sicherstellung der korrekten Übertragung des Funkverkehrs.

, Ende.	, à toi.
, ende.	, à toi.
, melde Dich.	, fais-moi signe.
, Melde Dich.	, fais-moi signe.
, Moment,	, un instant,
, moment,	, un instant,
, Moment.	, un instant.
, moment.	, un instant.
, richtig.	, c'est correct.
, Richtig.	, c'est correct.
, richtig.	, correct.
, Richtig.	, Correct.
, verstanden.	, compris.
, Verstanden.	, compris.
,moment.	un instant.
,Moment.	un instant.
,richtig.	, c'est correct.
,Richtig.	, c'est correct.

Abb. 5: Glossareinträge aus dem Bereich Funkverkehr

Und auch wenn der oben beschriebene notwendige Ausschluss von Homonymen im ersten Moment nachvollziehbar scheint, stellt die vermeintlich logische und eindeutige Anforderung den Nutzer in der praktischen Umsetzung

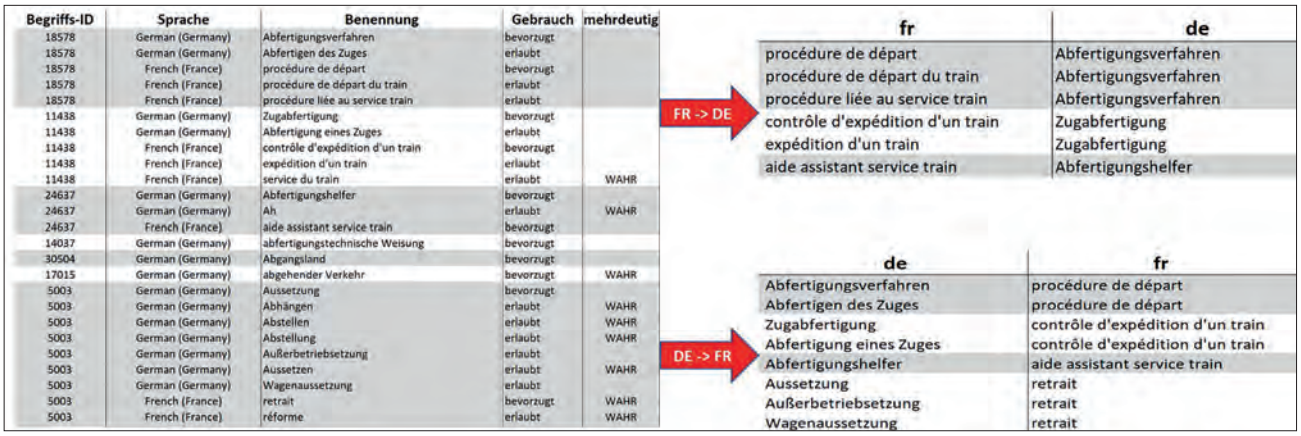


Abb. 4: Umwandlung begriffsbezogener Einträge in benennungsbezogene Glossare

immer wieder vor notwendige Abwägungen: „Bremse“ ist sicherlich ein Homonym, ebenso wie „Zug“. Aber sollte auf die Integration dieser für den Bahnkontext essentiellen Begriffe verzichtet werden? Oder wiegt die Erzwingung der im technischen Kontext korrekten Übersetzung die geringe Wahrscheinlichkeit der alternativen Bedeutung auf?

Beispiel: „Ein Schaffner steht im Zug“ oder „Ich trete auf die Bremse“

Es bleibt eine Ermessenssache im Graubereich eines jeden Projektes.

Evaluation

Auf Auswahl und Aufbereitung der Daten und anschließenden Trainings- oder Customization-Prozess folgt stets die Evaluation des NMT-Outputs, um den Erfolg der Beeinflussung zu überprüfen und im Zweifel Korrekturmaßnahmen einzuleiten. Dafür haben sich folgende Indikatoren und Maßnahmen bewährt:

BLEU

Der BLEU-Score ist trotz aller Ungenauigkeit und Fehlerhaftigkeit (basiert auf rein lexikalischem Vergleich des Outputs zu einer Humanübersetzung) weiterhin der meistgenutzte Qualitätsindikator. Zwar gibt es bereits bessere Ansätze, allerdings wird BLEU ob seiner großen Verbreitung und zum Zwecke der Vergleichbarkeit, auch von vielen Anbietern, weiterhin verwendet.

Automatisiertes Postediting (skriptbasierte Terminologieanalyse)

Zusätzlich zum BLEU-Wert empfiehlt sich eine (skriptbasierte) Kontrolle der Terminologieintegration. Dies ist recht einfach umzusetzen, indem alle Sätze, in deren Ausgangssprache eine Benennung aus dem integrierten Glossar enthalten ist, auf Vorhandensein des entsprechenden Äquivalentes in der Zielsprache überprüft werden. Die Auswertung liefert ein recht verbindliches Ergebnis des Integrationserfolges.

Humane Auswertung: Granularität der Fehlerkategorien und Kritikalität

Unabdingbar bleibt am Ende die humane Auswertung aller Fehler, kategorisiert in Fehlerkategorien. Die Granularität der Einteilung ist abhängig von den jeweiligen Projektanforderungen: Je höher die Qualitätsanforderungen sind, desto feingliedriger sollten die Kategorien sein. Im Bereich der Terminologie beispielsweise hat sich bewährt, die Fehler auf die Wortart oder den Grad der Flektiertheit hin zu analysieren.

Außerdem empfiehlt sich die Bewertung der Fehlerkritikalität, deren Matrix ebenfalls von Projekt zu Projekt variieren kann. In der Regel arbeitet die DB AG mit der Einteilung in „minor mistakes“, die die Verständlichkeit der Übersetzung nicht beeinflussen, „major mistakes“, die ein Nachfragen erforderlich machen, aber keine Gefahr darstellen, und „critical mistakes“, die eine Gefahr für Leib und Leben bedeuten können. Die Kritikalität ist stets Hinweis auf die Priorisierung der Fehlerbehebung.

Wie anfangs beschrieben, stellt der Kreislauf von Datenauswahl und Aufbereitung, Training bzw. Customization und der anschließenden Evaluation einen Kreislauf dar, der mehrere Zyklen beinhalten kann. Das sollte bei der Kosten- und Zeitplanung unbedingt berücksichtigt werden.

Use Cases der NMT bei der DB AG

An dieser Stelle seien zwei Use Cases der Neuronalen Maschinellen Übersetzung bei der DB AG vorgestellt:

DB Corporate Translate

DB Corporate Translate ist die datenschutzkonforme und datensichere NMT für alle Mitarbeiter im DB Konzern. Sie bietet Freitext- und Dokumentenübersetzung und wird via API in Konzernanwendungen integrierbar sein. Die Wahl der richtigen Form der Umsetzung war ob des divergenten Nutzerkreises eine Herausforderung: Hunderte Berufe aus verschiedenen Bereichen des Konzerns, teilweise mit unterschiedlicher Terminologieverwendung (z. B. Amerikanisches Englisch vs. Britisches Englisch) bedeuten eine hohe Anforderungsdiversität an Stil, Idiomatik und Terminologie der Übersetzung. Eigentlich ein klarer Fall für die korpusbasierte Schulung… Da klar war, dass mit einer Engine niemals alle Anforderungen abgedeckt werden könnten, wurde beschlossen, die NMT ausschließlich glossarbasiert mit der morphosyntaktischen Terminologieintegration zu optimieren.

Quelle für die Integration ist die vollständige zentrale Konzernterminologie über alle Sachgebiete hinweg, mit Ausnahme der Homonyme. Homonyme Benennungen werden bei der Datenpflege in der Terminologiedatenbank als „mehrdeutig“ gekennzeichnet und sind entsprechend filterbar. Die terminologischen Daten werden wie oben beschrieben für den Import pro Sprachpaar benennungsbezogen in Glossarform aufbereitet. Dieser Prozess funktioniert größtenteils skriptbasiert; allerdings sind noch viele manuelle Handgriffe notwendig. Obwohl die Terminologiedatenbank täglich wächst und aktualisiert wird, findet die Integration nur in Intervallen von mehreren Wochen statt.

Im Web-Interface bietet die Konzern-MT eine Terminologieerkennung, bei der in einem Mouseover-Fenster die wichtigsten Begriffsinformationen angezeigt werden (z. B. Definition und Hinweis zum Gebrauch). Via Link zur zentralen Terminologiedatenbank sind alle weiteren Informationen schnell griffbereit.

Abb. 6: DB-interne NMT (DB Corporate Translate) mit integrierter Konzern-Terminologie

Bereiche, Abteilungen und/oder Projekte der Bahn mit spezifischeren Anforderungen haben die Möglichkeit, sich eigene Engines trainieren zu lassen. Ein Beispiel dafür ist das im Folgenden beschriebene Projekt.

KITT - KI Translation Tool für die Kommunikation zwischen Lokführer und Fahrdienstleiter

Spätestens aus einem Unglück heraus entstand die Idee, auch den grenzübergreifenden Zugverkehr sprachtechnologisch zu unterstützen: Als sich im August des Jahres 2017 auf Höhe Rastatt das Gleisbett der Rheintalbahn absenkte, blieb die wichtige Nord-Süd-Strecke über Monate blockiert. Der Güterverkehr staute sich, und in deutschen, niederländischen und italienischen Häfen stapelten sich die Container. Zwar existieren Ausweichstrecken über Frankreich, aber aufgrund der Ferienlage standen keine französischsprachigen Lokführer für die zusätzlichen Züge zur Verfügung. Neben dem Triebfahrzeugführerschein benötigen Lokführer nämlich auch Streckenkenntnis sowie Sprachkompetenz für alle durchfahrenen Länder auf mindestens B1-Niveau. Grund also für die DB AG, sich langfristig durch technische Mittel von diesen Einschränkungen zu befreien.

Die Lösung ist bereits seit Längerem in Arbeit und besteht im Wesentlichen aus einer Reihung dreier Komponenten: Spracherkennung (Speech To Text, STT), NMT und synthetische Sprachausgabe (Text To Speech, TTS). Auch die Behandlung sicherheitsspezifischer, standardisierter Mitteilungen im Bahnverkehr (predefined messages, PDMs) ist sichergestellt, was an dieser Stelle aber nicht weiter thematisiert werden soll. Grundsätzlich funktioniert die Anwendung also wie folgt: Die Spracherkennung transkribiert das „Verstandene“ automatisch, die NMT überträgt den Transkripttext automatisch in die Zielsprache und die Sprachausgabe gibt ihn mittels synthetisierter Form aus.

Jede Komponente an sich hat ob ihrer KI-Basiertheit ein Risiko inne, welches sich durch die Verkettung der Komponenten entsprechend potenziert. Das verlangt eine akribische Feinabstimmung an jeder Komponente sowie an den Schnittstellen. Typische Fehlerquellen der jeweiligen Komponenten sind unter anderem:

STT:

- Interpunktion (keine/fehlerhafte Kommata, keine Ausrufezeichen, keine Fragezeichen),
- Homophonie („Eisläufer“ statt „Heißläufer“), Akronyme und Code („Tee“ statt „T“),
- Geräuschquellen (Zugmotor),
- Sprechereigenschaften (Dialekt, Akzent, subjektive Stimmcharakteristik)

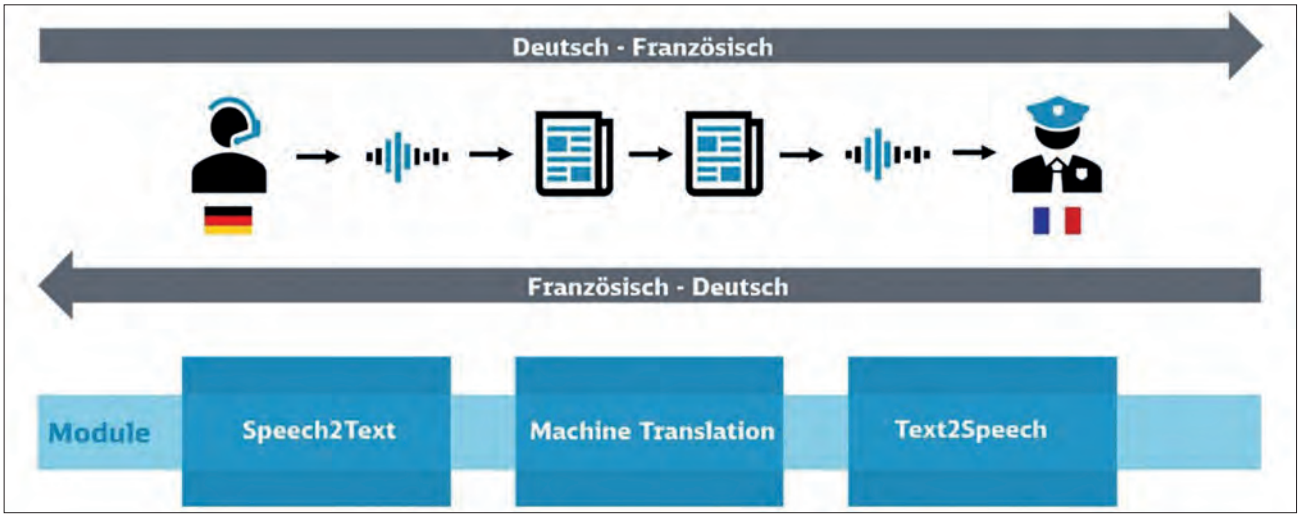


Abb. 7: Komponentenkettung STT – MT – TTS

NMT:

- Terminologie (Ablaufberg, Schotterzwerg),
- Idiomatik („Person im Gleis“ -> „Personnes sur les voies“),
- Stil (Umgangssprache: „Kannst‘ weiterfahren“),
- Pivotsprache („Conducteur de train où est-ce que tu es hétéro?“)

TTS:

- Ausgangssprachlizismen im Zieltext (Bahnhofsnamen, Aussprache von „ICE“ als „Eiss“),
- Geschwindigkeit

Eine besondere Herausforderung dieses Projektes besteht in der sehr spezifischen Sprache in Bezug auf Stil, Idiomatik und Terminologie: Die Kommunikation zwischen Lokführer und Fahrdienstleiter ist äußerst umgangssprachlich („Kannst‘ losfahren.“), weicht stark vom normalen Sprachgebrauch ab („Person im Gleis“, „Ich stell‘ Dir das Signal.“) und ist voller Fachausdrücke („Hab‘ nen Heißläufer.“, „Schotterzwerg“ etc.).

Aufgrund dieses extrem speziellen Anforderungsprofils bleibt als NMT-Beeinflussung nur die korpusbasierte Form, um die NMT auf Stil und Idiomatik sowie auf Fehler bzw. Besonderheiten der jeweils vorangehenden Komponente zu sensibilisieren. Dafür wurden verschiedene Anbieter und verschiedene Beeinflussungsformen getestet: korpusbasierte Customization, korpus- und glossarbasierte Customization, korpusbasiertes Training.

Bei aller Anspruchsvarietät der NMT-Anbieter in Bezug auf Menge und Aufbereitungsform des Trainingskorpus hatte die oberste Regel stets Bestand: Trainingskorpus und Glossar müssen in der Ausgangssprache variantenreich, in der Zielsprache sehr konsistent gestaltet sein, um die Outputqualität zu steigern.

Die Qualität wurde und wird weiterhin in regelmäßigen Abständen sehr granular evaluiert: Terminologiefehler werden beispielsweise nach Wortart und Wortanzahl getrennt betrachtet. Die dadurch aufgedeckten kategorischen Schwächen bei der Integration von Verbalphrasen und stark flektierten Wortarten führten schließlich zu einer Art projektintrinsischer Evaluation der Terminologieintegration: zunächst nur korpusbasierte Customization, dann korpus- und glossarbasierte Customization (s. Manipulation), schließlich die Erweiterung des Glossars um flektierte Formen (erweiterte Manipulation), hin zum rein korpusbasierten Training: der Provokation.

Die morphosyntaktische Terminologieintegration ist für dieses Projekt weiterhin sehr wünschenswert, allerdings erlaubt der einzige Anbieter dieser Lösung kein korpusbasiertes Customizing. Damit ließen sich zwar die terminologischen Projektspezifika umsetzen, jedoch zulasten der Anforderungen an Stil und Idiomatik…

Fazit

Wie eingangs beschrieben ist die NMT-Beeinflussung derzeit ein großes Experimentierfeld von Anbietern, Systemen und Funktionen, deren Einsatz und Nutzen im Fokus der jeweiligen Projektanforderungen analysiert werden müssen. Aufkeimende Entwicklungen, wie die morphosyntaktische Terminologieintegration von DeepL, versprechen weitere essenzielle Entwicklungsschritte in den kommenden Jahren, die diese Analyse immer wieder aufs Neue notwendig machen. Aktuell existiert allerdings auf dem Markt keine Lösung, die die Kombination aus korpusbasierter Customization und morphosyntaktischer Terminologieerkennung und -integration erlaubt. Zunächst kommt es also weiterhin auf den Kontext an.

Literaturverzeichnis:

[1] DB Language Portal, die Terminologiedatenbank der Deutschen Bahn AG. (https://dblanguageportal.noncd.db.de/Lookup.aspx) [Zugriff am 25.09.2021]

[2] Gleisabsenkung bei Rastatt. (https://bahnblogstelle.net/2017/08/16/nach-gleisabsenkung-bei-rastatt-rheintalbahn-bleibt-noch-wochenlang-gesperrt/) [Zugriff am 21.09.2021]

[3] Interview mit Samuel Läubli (2021): „Maskierung und Provokation. Wie sich die Terminologiequalität bei NMT verbessern lässt“ In: edition 01/2021, S. 38-43.

[4] KITT Translation Tool. (https://www.youtube.com/watch?v=fmjLsJ1MvBM) [Zugriff am 29.09.2021]

[5] METEOR Automatic Machine Translation Evaluation System. (https://www.cs.cmu.edu/~alavie/METEOR/) [Zugriff am 22.09.2021]

[6] Nicole Keller (2021): „DeepL integriert Glossarfunktion“ In: edition 01/2021, S. 34-35.

[7] Tom Winter, Daniel Zielinski (2020): „Terminologie in der neuronalen maschinellen Übersetzung“ In: Maschinelle Übersetzung für Übersetzungsprofis, S. 210-233.

Tom Winter arbeitet als Terminologe, Computerlinguist und Data Scientist im Sprachmanagement bei der DB AG. Neben dem Terminologie- und Sprachdatenmanagement liegt sein Schwerpunkt auf der Integration, Beeinflussung und qualitativen Bewertung von MÜ-Systemen.

Kontaktadresse
Tom.Winter@deutschebahn.com
www.deutschebahn.com

Quality Evaluation of Termbases

Beate Früh and Dóra Mária Tamás

The paper aims to present an experimental set of criteria useful for the evaluation of the quality of termbases. The main areas addressed are the environment, the technical parameters, the information on the content and usage of termbases. As part of the quality evaluation process, the concept of data validation is explained. For a practical approach, the article includes some examples as well as information on data maintenance and shows quality assurance features of some commercial translation management systems.

Keywords: quality evaluation of termbases, quality criteria catalogue, terminological entry, terminology management tools, data validation

1. Introduction

To make best use of a termbase system, it is necessary that the user understands the structure and methodology of the software. During the course of many years of development, termbase systems have become more complex to meet technical and professional requirements – despite the fact that the first termbases of large organisations were created or were already operational in the 1980s (see TERMIUM Plus of the Translation Bureau of Canada, Euskalterm of the basque UZEI Centre and TERMDAT of the Swiss Chancellery). Modern terminology and its tools are relatively new compared to the area of lexicography, which of course boasts of a far longer tradition. Therefore, there is still a gap in terms of developing standardised and objective evaluation criteria for termbases. Based on Sager (1990), the first generation of termbases in the early 1970s were built in general on ad hoc data management and on a lexical rather than on a conceptual basis, though Tanke (2008) states that the first electronic termbase was already operational in 1969 and concept-oriented.

Termbases can be created for different purposes (e.g. language policy, translation objectives, standardisation) by different types of organisations (e.g. translation companies, universities, research institutes, international organisations or corporations of the industrial and business area) and can contain concepts of different subject fields (e.g. legal, economic, medical, technical, etc.). Despite showing common features, it is easy to recognise that there are differences. Therefore, it is not easy to create a generally accepted evaluation system that covers all possible aspects. This paper aims to fill this gap by offering a set of criteria applicable not only for online termbases but also for the description, checking and auditing of internal termbases. An audit of a termbase usually consists of a report, comments on non-conformities and recommendations for improvement. But which aspects should be examined and on which principles and methods should the examination be based?

Our set of criteria presented in the paper was elaborated based on technical literature, standards and practical examples of termbases. As a starting point, we have taken into consideration the evaluation criteria developed for printed and electronic dictionaries (Ripfel 1989, Hartmann 2001, Fóris/Rihmer 2007, Gaál 2012) and the formal evaluation of central online termbases worked out by Tamás/Sermann (2019), who limited their research to online surfaces only. Finally, we developed our own criteria catalogue. Compared to Tamás/Sermann (2019), we have analysed not only external termbases of larger organisations, but also internal termbases. We have also taken into account the more practical approaches of terminology management (e.g. Frey/Schmacht 2010, COTSOES 2018, Drewer et al. 2014, Drewer/Schmitz 2017, Schmitz 2020). Of course, we did not intend to ignore standards, which are often more focused on specific features (e.g. terminology work and harmonisation in ISO 1087 (2019), data categories in ISO 12620 (2019), the indication of bibliographical references in ISO 12615 (2004), data maintenance and updating in ISO 12616 (2002) or ISO 26162-1 (2019) about the management of terminology resources) and less on describing a comprehensive system, though the standard ISO 23185 (2009) also addresses a wide range of aspects, but without describing all details of the given requirement.

2. Basic concepts and relevant aspects of quality evaluation and terminology management

The interpretation and definition of quality poses several difficulties, since quality always depends on a specific field or environment. In our paper, we have interpreted quality in terminology management as the consistent and appropriate application of widely adopted and professionally acknowledged principles and methods of terminology work and documentation. The terminology work and the termbase content vary to a certain extent from organisation to organisation based on different considerations (e.g. purpose of the termbase, target group, domain and concepts). Therefore, the quality of termbases can be examined under various aspects.

2.1 Definitions

To start with, we need to define some basic concepts used in this paper. When we refer to a **termbase** (or terminology database), we understand this as the collection of electronically stored terminological data created from a concept-oriented approach and based on mapping the conceptual system of the subject field, which contains terms and their definitions of one or several subject fields in one or more languages (Tamás/Sermann 2019). We do not differentiate between termbase and term bank (terminological data bank), the latter of which according to ISO 1087 (2019) is a collection of multi-lingual terminology databases, which includes the organisational framework for recording, processing and disseminating terminological data.

It is necessary to clarify what we understand by the following key concepts: quality assessment, quality assurance, quality control and quality evaluation.

The ISO 23185 (2009) standard defines assessment as a "process to demonstrate that a terminological resource fulfils specified requirements" though we have been more focused on the evaluation of quality as a result and less on the evaluation process. In our interpretation **quality assessment** refers to the analysis of the technical and interpersonal aspects of a terminology process and its outcomes, and in addition how to control and improve it. **Quality assurance (QA)** is a proactive process to prevent quality nonconformity of a terminological product. QA reveals and fixes the sources of possible quality problems. QA measures already begin with the development of the termbase and usually end before or during the entry of a concept into the termbase (see Schmitz 2020). Quality assurance deals with establishing guidelines, training for terminologists and users and designing processes.

In contrast, **quality control (QC)** is a reactive process to detect quality nonconformity of a terminology product e.g. the content of a termbase. **Quality evaluation** is both – assessing the QA measures taken to ensure the data quality of the content of the object to be evaluated (e.g. a termbase) and performing quality checks of the termbase to look for errors not yet fixed to gain an insight into the quality of the termbase content. The result is a rating of the assessed termbase and an expert report containing a catalogue of recommended measures that should be taken to fix the detected quality deficiencies (nonconformity). Recommendations are also provided for necessary, additional QA measures to avoid the already identified quality nonconformity in the future.

2.2 Use cases for quality assessment

After having defined some basic concepts relating to quality and quality assessment, we would like to give some information on use cases and on when it is helpful or even necessary to assess the quality of a termbase. We have identified the following cases:

- No database maintenance has taken place for a long time.
- Many different persons have contributed to the content, but no common ruleset has been established and inconsistencies and double entries are the result.
- Outdated terminology has been reported and calls for action.
- An external termbase needs to be assessed to determine if content complies to one's own standards and if the termbase can be used in addition to the organisation's own resources.
- A merger of two organisations takes place and it is required to check if actions need to be taken regarding the terminology of the "new organisation".
- Terminology referring to a new subject field of business needs to be added.
- New languages need to be added.
- The migration of the termbase to a new terminology management system or a modern knowledge organisation system (KOS) is planned.

3. Evaluation system

Looking at a large set of different resources and our own experience, we have put together a large criteria catalogue for the evaluation of termbases (see Fig. 1). This evaluation system revolves around the following four main criteria:

I. environment,
II. technical parameters,
III. structure and content,
IV. usability and features of the termbase.

Fig. 1: Quality evaluation of termbases – upper levels of the criteria catalogue (own figure). The mind map showing the complete criteria catalogue may be downloaded using the following link: https://www.buerob3.de/service/

Before starting to evaluate a termbase, the framework is to be set. The first two main categories help to build a frame. Due to their extent, the paper does not present a use case that includes all criteria (although there are case studies of analyses based on Tamás/Sermann (2019), see also Tamás (2021) about WIPO Pearl and the set of criteria that is in use in workshops and teaching in higher education in Hungary).

When we were preparing this evaluation, we encountered the same difficulties which are described in ISO 23185: "Although in the following clauses the attributes are described individually, it is necessary to bear in mind that the usability attributes can be interrelated with, and [are] dependent on each other" (2009:4). Tamás/Sermann (2019:35) state similarly: "features are closely related to each other, and often hard to separate (such as editing principles, documentation, reliability and quality)" and "some features belong to more than one criterion".

3.1 Environment

Environment is the first main category of our evaluation system and is divided into two main groups:

- the organisational history and
- the specific environment of termbases.

The first main group, the organisational environment, lists the history of termbases and tendencies as subcategories. The subcategory of the history of termbases includes the historical background and the relationship with other termbases or databases. The historical background, which may reveal if the termbase has a predecessor, as well as the date of creation of the current version, help in order to gain a picture of the evolution of the termbase. So it becomes clear whether we have to evaluate a newly implemented or an improved version of a former database; although the evaluation is first of all focused on the current state and thereby on the synchronic analysis of the termbase, which describes its current features. However, a diachronic description may be useful in certain cases (i.e. if it has a predecessor like IATE). Looking at the relation of a termbase with other termbases or databases within or outside the organisation can also provide useful and relevant information (e.g. the termbase IATE is connected with the legislative directory of EUR-Lex and has various subcentres of edition, the terminology portal of WIPO Pearl, the Patentscope database containing patents). The subcategory of tendencies within the organisational environment focuses on the fact whether the termbase has been created to fulfil specific objectives such as:

- terminology policy aims (e.g. the support of a language minority such as in the case of the termbase bistro);
- translation orientation (e.g. IATE, which serves primarily to provide support for translators and interpreters in the translation of EU-administrative texts);
- standardisation purposes (e.g. corporate language).

The second main group of the main category of environment, namely "specific environment", has three branches: characteristics of the termbase provider, the target users and the type of terminology work. As for the characteristics of the termbase provider, we can distinguish for instance international organisations (e.g. WIPO Pearl), public administrative bodies (e.g. TERMDAT), research institutes (e.g. bistro), translation agencies (e.g. TERMIUM Plus) or companies (e.g. SAPTerm). The second subcategory of target users includes native speakers (having a majority or minority status), translators, experts of a specific domain (e.g. engineers, lawyers, physicians, etc.), employees working in different areas like marketing or technical documentation and terminologists.

By examining the third subcategory, namely the type of terminology work contained in the termbase, we can distinguish between

- descriptive or prescriptive terminology work,
- monolingual and multilingual terminology work or
- a systematic (domain-oriented, text-oriented) or ad hoc terminology work.

3.2 Technical parameters of a termbase

The second category of our evaluation catalogue deals with the technical parameters of the software. The first check is whether the termbase is self-developed by the organisation publishing the termbase or whether it is a commercial tool or a hybrid solution. Next, we assess the level of accessibility,

16./17.09.2022
Berlin (Kosmos)

Bundesverband der
Dolmetscher und Übersetzer

BDÜ

7. FACHKONFERENZ SPRACHE UND RECHT

Rechtsvergleichung in der Praxis

Am 16./17. September 2022 veranstaltet der Bundesverband der Dolmetscher und Übersetzer e. V. (BDÜ) die siebte Auflage seiner Fachkonferenz zum Thema „Sprache und Recht", diesmal unter dem Motto: „Rechtsvergleichung in der Praxis". Die Konferenz richtet sich an Dolmetscher und Übersetzer, die mit Recht und Rechtssprache arbeiten, sowie an interessierte angrenzende Berufsgruppen wie Rechtsanwälte, Richter, Notare, Vertreter der Polizei und Justiz und Vertreter von Behörden. Ihnen allen bietet die Konferenz eine Diskussions-, Weiterbildungs- und Netzwerkveranstaltung rund um die Themengebiete Recht in verschiedenen Sprachen und Systemen, Rechtssprache, Rechtsvergleichung und praktische Arbeit der Rechtsdolmetscher/-übersetzer. Erwartet werden rund 500 Teilnehmer aus Deutschland und dem Rest der Welt sowie Delegationen von Berufsverbänden der Dolmetscher und Übersetzer aus zahlreichen Mitgliedstaaten der Europäischen Union.

Anmeldestart: 1. Dezember 2021 – weitere Informationen und Anmeldung unter:

→ www.fsr2022.de

which depends much on the focus of the termbase, and its target user groups. Does it offer a public access without needing any credentials, or is the access restricted? Perhaps, only certain data categories are publicly accessible, or domains are with restricted access. Another question is whether download options are granted. Another interesting aspect of accessibility is if the content is for example based on collaborative work such as crowdsourcing and how access is granted in this case. There might be even termbases where part of the content may be accessed via different platforms. In times of smartphones and other mobile electronic devices, the accessibility to terminology data via those tools is, of course, gaining more and more importance. Modern terminology work tends to be more collaborative teminology work, and so an increasing number of tools offer features such as commenting and workflows. Another important technical aspect is the exchangeability of data (e.g. import, export formats and interfaces). More details are described in section 3.4 Usability and features of the termbase.

3.3 Structure and content of the termbase

With regard to the structure of a termbase, ISO standards started to set internationally recognised requirements for professional terminology management long ago (see ISO 12620 Management of terminology resources – Data category specifications with versions of 1999, 2009 and 2019; ISO 12616 (2002) Translation-oriented terminography; ISO 16642 (2017) Computer applications in terminology – Terminological markup framework, ISO 26162-1 (2019) Management of terminology resources – Terminology databases – Part 1: Design).

General features

Modern terminology work requires formal criteria of the database structure such as concept orientation and term autonomy. One possibility to classify a termbase is – according to Tamás/Sermann (2019:27-28; 43), who distinguished between "simple, traditional or complex termbases" (see Table 1) – based on the complexity of their structure. As the table describes, it may be important for certain organisations whether the content is created on the basis of a purely onomasiological, i.e. concept-oriented approach or whether the content is based on a hybrid approach containing semasiological elements typical for lexicography as well. But a concept-oriented approach is mandatory in order to define these tools as terminological databases. Certain structural elements show whether a termbase is rather a traditional one or a modern, knowledge-oriented termbase with corpus (e.g. laws in case of a term bank about legal concepts, see EOHSTerm) or it shows ontological elements like concept maps (see WIPO Pearl).

Modern, knowledge-oriented termbases include elements from knowledge organisation systems (KOS) like the ability to create taxonomies (see Coreon), which could be added as a further aspect for extra elements.

Detailed structure

Evaluating the termbase structure in a more detailed way, we distinguish between four categories, namely:

- the megastructure,
- the macrostructure,
- the microstructure and
- the mesostructure.

Those four categories are traditionally used in lexicography, whereby apart from Tamás/Sermann (2019) terminology refers to macrostructure and microstructure in ISO 1087 (2019) Terminology work and terminology science – Vocabulary. At the level corresponding to megastructure of vocabularies, criteria like the availability of directions for use and other modern forms of help (forum, chat service for users facing difficulties, copyright) are considered. The availability of key figures like information on the number of languages (monolingual, bilingual or multilingual), concepts/entries and terms or number of domains and subdomains form also part of the **megastructure**.

Type of termbase	Concept-oriented	Term autonomy	Number of data fields and extra elements
SIMPLE	mandatory (or with some hybrid features)	not necessarily applicable (resembles a word list)	their number does not reach the minimum (e.g. only term and subject fields)
TRADITIONAL	mandatory	applicable	traditional number of data fields present, definition in at least one language
COMPLEX (sometimes called knowledge base or terminology information system)	mandatory	applicable	traditional number of data fields present and further pieces of knowledge (concept maps, corpora)

Table 1: A tentative classification of termbases (based on Tamás/Sermann 2019:28)

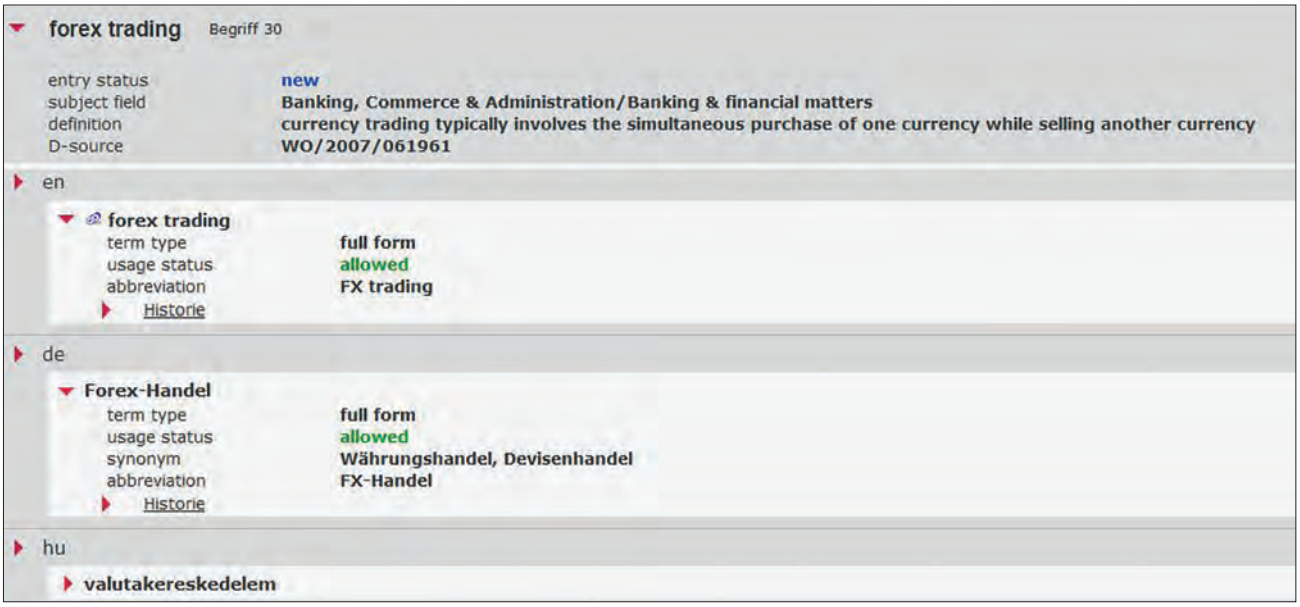


Fig. 2: Term entry not compliant with term autonomy (own figure)

Looking at the level corresponding to **macrostructure** this category includes different search options starting from simple search, domain- or sub-domain-specific search, full-text search or search for text fragments reaching to advanced search features for terms like right-hand and left-hand truncation, Boolean operators, and other filters (see also ISO 12616 (2002)). Search options can be extended to an aligned corpus (EOHSTerm), to the use of links giving access to other data repositories like EUR-Lex in the case of IATE or Patentscope in the case of WIPO Pearl, to a hit list helping to retrieve information and to the nature of information shown (e.g. text-based or images and other types of multimedia).

Whereas, when referring to the **microstructure** of the termbase, we are looking in more detail at the data categories used in the termbase. As Tamás (2021) states at the level of microstructure, the number of data categories is a clear indicator of the extent of detail in the database. It can be seen whether the termbase contains minimal information or whether it is a well-developed, detailed data repository; and this information is grounds for evaluation for the user.

The display of data categories and the clarity of entries expressed by the order and labelling of data categories can also be a quality indicator. The more data categories the entries contain, the greater is the information value of the database and the more diverse its usability. However, as the number of categories increases, the entries become less manageable, and the maintenance of the database becomes more and more time-consuming (see COTSOES 2018). Too large a variety of data makes it even more difficult to query the database or extends the query time. Entries that are too detailed or over-queried run the risk of not being read in their entirety by the language experts, who are often working under time pressure. In the worst case, wrong or inappropriate equivalents are selected because the domain or subject field, the definition, or a note on usage is overlooked. Therefore, it is necessary to find a reasonable middle way.

A very important aspect is looking at the adherence to the following basic principles of terminology data modelling as laid down in ISO 26162-1 (2019). Apart from the must of **concept orientation**, based on which a terminological entry (or record) should refer to one concept, **term autonomy** (see Fig. 2) should also be applied, a principle based on which "all terms in a concept entry are considered independent sub-units and can be described using the same set of data categories" (ISO 26162-1 (2019:4)). Fig. 2 illustrates the non-compliance with term autonomy, as abbreviations and synonyms have own data categories on term level and are not processed as autonomous terms within the entry. The level of implementation of a three-level **hierarchical structure** (entry level – index level – term level) is often already system-dependent and not always transparent through external descriptions, i.e. self-developed online surfaces. There are two more content-related quality indicators referring to the **data elementarity** (a data category contains only one element) and **granularity** (level of precision).

If we look at ISO standards on preparing termbases, we can find recommendations specifically about the data categories in the already mentioned ISO 12620 (2019) standard titled "Management of terminology resources – Data category specifications". Examining data categories from

a structural point of view, it might also be helpful to see whether fixed fields or additional customisable fields are available. Most of the tools today offer the option to label data fields as mandatory or optional. One data category needs to be assessed in a different way: the definition. “Owing to the concept-oriented nature of a termbase, the definition undoubtedly plays a central role beyond the term, while equivalence is not always a separate category” (Tamás/Sermann 2019:34-35). Conceptual equivalence information is often contained in the definition or a separate note. Another quality indicator may be the number of languages of the definition. Besides the aspect that the definition can be substituted by a context, a critical look at the formal presentation of definitions should be given (e.g. ensure proper intensional definitions, see ISO 704 (2009)).

Speaking of context, it may be worthwhile to check if a policy has been laid down regarding the preference for defining, or if associate or explanatory context sentences shall precede over definitions. Can context information or other context-related metadata such as the domain information replace missing definitions? The question is not an easy one to answer. The traditional function of a context is to offer information about the usage of a term in a brief text sample. By replacing definitions with contexts, extra information is quickly added and costs less than a definition worked out by an expert, but it is less precise by indicating only a few properties of the concept.

Another data category very closely related to the concept is the domain classification. As to Kudashev (2013:20) “domain classification is of the utmost importance for term banks as they typically contain terminology from multiple subject fields”. In some termbases, concepts can be assigned only to one domain, whereas other tools allow multiple domain classification.

Depending on the nature of the termbase, additional fields on the term level such as the processing status, the term type and usage or acceptance of a term (recommended, forbidden, standardised, obsolete, see also the DatCatInfo data category repository for language resources) are necessary. Special labels for synonyms and a policy how to treat quasi-synonyms are optional. The handling of lexical synonyms, short forms and “cosmetic inconsistencies” due to different spelling or form (e.g. checkbox vs. check box) “although they seem minor, can be as equally damaging to the leveragability of translation memories as lexical synonyms” (Warburton 2015:328). “Inconsistencies of a lexical nature (lexical synonyms) in the source language are likely to be repeated in target languages.” (ibid.) If term checker tools for authoring purposes are fed by the termbase, the termbase needs to match the requirement by collecting synonyms, and labelling them with administrative status as preferred, allowed and forbidden term, so that those tools can fulfil their QA purposes.

To state the reliability of a term or concept, there are termbases that contain information whether the usage was confirmed by a terminology committee (see for example the termbase bistro). Another measure to state the reliability and validation status is to use a quality label (see Sager 1990: 141). For more details see "reliability" in section 3.4 Usability and features of the termbase.

At the level corresponding to **mesostructure**, technical aspects such as cross-references (links within the entries, links among entries) are checked. External links to other websites or repositories must be checked for their validity and their appearance (can they be opened and accessed).

3.4 Usability and features of the termbase

The main category of usability and features includes as subcategories user-friendliness, updates, reliability of data, innovative nature and last but not least the social value and professional importance of the termbase.

User-friendliness plays a key role in the everyday use of a tool, because it will determine how many searches will be undertaken and as a consequence whether the organisation will really benefit from the existence of the termbase. For this reason, it is necessary to take into account whether the user interface can be easily understood, whether it is worth using abbreviations (like in an earlier version of TermDat), which are the visual elements that facilitate the search and the reading of data. For instance, a customisable data structure with flexible data fields can also contribute to a more frequent use, since the information can be retrieved more quickly (e.g. the accessibility to a full entry or to all domains or single domains in bistro and WIPO Pearl).

The exchangeability of data can represent an important aspect as well as already mentioned in 3.2, since the compatibility with other data structures allows the easy fusion with other datasets. We agree with Schmitz (2020) who states that the lack of a thorough planning not only makes data exchange more cumbersome, but is later, once a termbase has been filled with data, more difficult and costly to correct and hinders communication with other systems, i.e. is an obstacle to interoperability.

Updates include two main features, this means the general state of up-to-dateness (dates of recording) and the frequency of updates. “Managing a database requires regular data maintenance and updating. This means revising, supplementing, or correcting existing entries on the basis of new material” (ISO 12616 (2002:8)). Update information should be clear

and visible not only for the editors (terminologists), but for the users as well, and updates should be carried out regularly (see DrewerSchmitz 2017). The regular database maintenance is also emphasized by COTSOES (2018), who sees it as an ongoing task. According to COTSOES (2018) this includes not only updating the content, but also adding to it, deleting multiple entries (duplicates), correcting duplicate and incorrect entries, as well as the formal cleaning, correction and adapting of information (e.g. to new spelling rules). An indispensable prerequisite for such tasks to be carried out is a database structure that allows to filter by metadata fields (status fields, subject areas, term types, administrative data). Particularly, in the case of large databases, data maintenance cannot be done on an ad hoc basis, but must be organised in projects and have – at its best – a fixed place in work planning based on the availability of necessary human and/or financial resources.

Another prerequisite for successful database maintenance is the continuous monitoring of terminological developments in the subject fields covered and in the specialist literature. Terminology managers must monitor these developments in the areas relevant to their terminology database together with the experts. This is the only way to ensure that the necessary updating of the database can be achieved in good time and reliable quality. It should be imperative that terminology managers establish a kind of "data maintenance policy", this means determine the priorities and criteria by which the data maintenance of their data collections will be planned and carried out.

The third subcategory indicating the reliability of data consists of different areas to be examined, such as documentation in general with the existence of a guideline with pre-defined editing instructions (e.g. mandatory and optional data categories, data validation conditions) and with the proper indication of the sources (e.g. bibliographic references in original language, internal or external contributors such as subject matter experts, use of permanent or one-off links). There is no doubt at all that a consistent source indication is a very important feature for traceability and reliability. ISO 12615 (2004) recommends that content, form and structure of bibliographic references shall preferably be in accordance with ISO 690 (2021), but the authors also think that own guidelines can serve this purpose as well.

Some databases have a so-called reliability index, an indication of the validation level of an entry and its terms. This index supports users in selecting terms with more or less precision.

Another criterion for high reliability is that the subject matter expert or a network of experts is involved in validation. Low(er) reliability is given when a data collection is the result of crowdsourcing. Successful collaborative terminology work is ensured if some preliminary measures are taken. In certain cases, it might also be recommendable to introduce a system of rewards and other incentives for completing the introductory terminology training and helping other users.

Grounded on the above, the main category of reliability is the area of documentation, the elaboration and presentation of data, the type of terminology work and subject matter contribution. In fact, reliability can be considered the core of the terminology work and the quality of the terminological data. It affects all areas, and is a category interrelated with several features.

The innovative nature of the TMS includes extra elements (e.g. concept map, corpus), extra information (e.g. video tutorial, e-learning, chat service) and an innovative display. There are termbases with the option of access from different platforms such as smartphones or tablets (see "Technical parameters" in section 3.2 and “A brief insight into terminology tools” in section 4.2).

Termbases as means of knowledge transfer can represent a significant value for the society, and therefore we have included as the last subcategory the social value and professional importance of termbases. It explores the value of the termbase from the perspective of terminology policy, the scope and type of experts interested (target users), whether the termbase producer offers opportunities for training, e-learning, workshop, cooperation agreements and university projects.

4. Data validation and tool support

In the preceding sections, we concentrated more on the technical and structural features of a termbase, in the following we will point out, based on some practical examples, how data validation can be supported by terminology tools. According to ISO 1087-1 (2019), data validation is “carried out to determine if terminological data are correct, complete or meet specific criteria”. It is therefore a core process of quality assurance, but only part of it can be carried out with tool assistance, a lot of validation checks still depend on the competence of an experienced terminologist

4.1 Data validation

Data validation is a quality control method to ensure that terminological data are formally and linguistically correct and consistent and that the new record is not a duplicate entry (see also Fig. 3).

Schmitz et al. (2020:M2-19) states that “The quality should be checked before it is approved. […] Validation focuses on adherence to best practices and methodologies […] and to

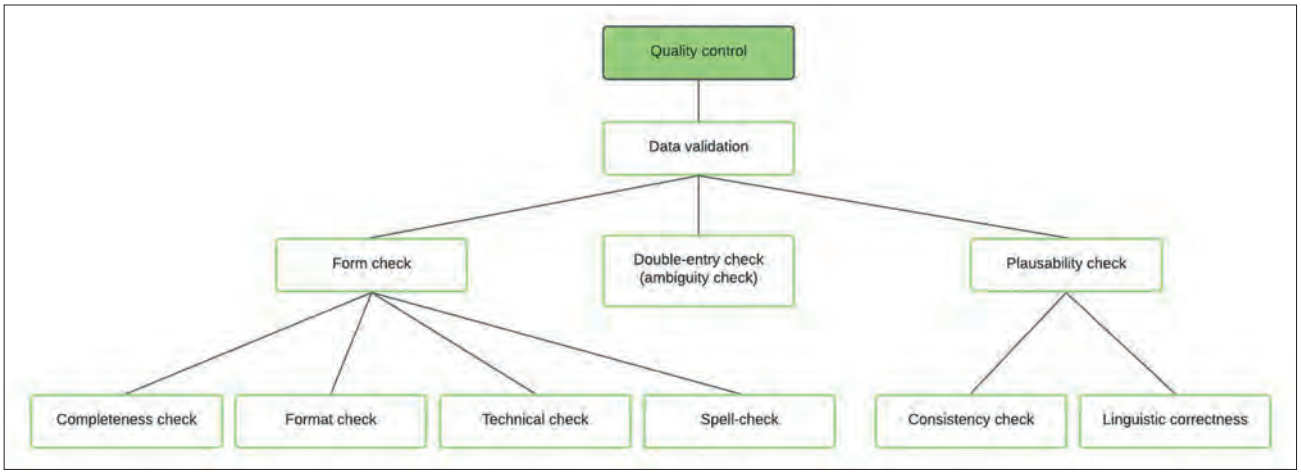


Fig. 3: A summary of the different validation checks (own figure based on ISO 1087 (2019))

the guidelines set by company or organization. […] Some validation aspects pertain to the entire termbase content, some pertain to individual entries. Validation can focus on content, methodology, linguistic or technical aspects." ISO 23185 (2009:11) states that "Before any data are accepted to be recorded in a terminological resource, necessary validation shall be done to ensure that the input data fit the overall requirements of the terminological resource."

When such validation has taken place, the status can be indicated by using the already mentioned reliability index, which is an information on reliability in each termbase entry and indicates by symbols, by icons, or by text and figures the reliability for the validated state.

Figure 3 shows the different types of data validation according to ISO 1087 (2019) though there are also other types of classifications (see Schmitz et al. 2020:M2-19). The three main areas covered are: formal check, double-entry check and plausibility check. **Formal checks** such as the spell check are a data validation carried out to determine that all words comply with predefined spelling rules. The formal check is also to check whether methodology has been followed and conventions as established in the organisation's terminology guidelines are followed. Are terms recorded in canonical form? Have all languages been provided? Is the term well-formed and following the term formation rules? The **technical check** is a specific form of the formal check and refers to technical aspects such as correct and functioning cross-references and correct display of special characters. The **completeness check** is carried out to see whether terminological data are present where required. Especially if all mandatory information is available and correctly provided.

The **double-entry check** (see Drewer/Schmitz 2017) (also known as duplicate control or homonymy control or latest named ambiguity check) ranks on a high level. The double-entry check is what ISO 23185 (2009) refers to as control of redundancy, e.g. control and avoid identical data content in a terminological resource. COTSOES (2018) states that double entries create noise and time loss during database queries, which is why they are unpopular with users. Therefore, they should be avoided as far as possible when creating new entries (see Fig. 4 as an example of double-check control in crossTerm).

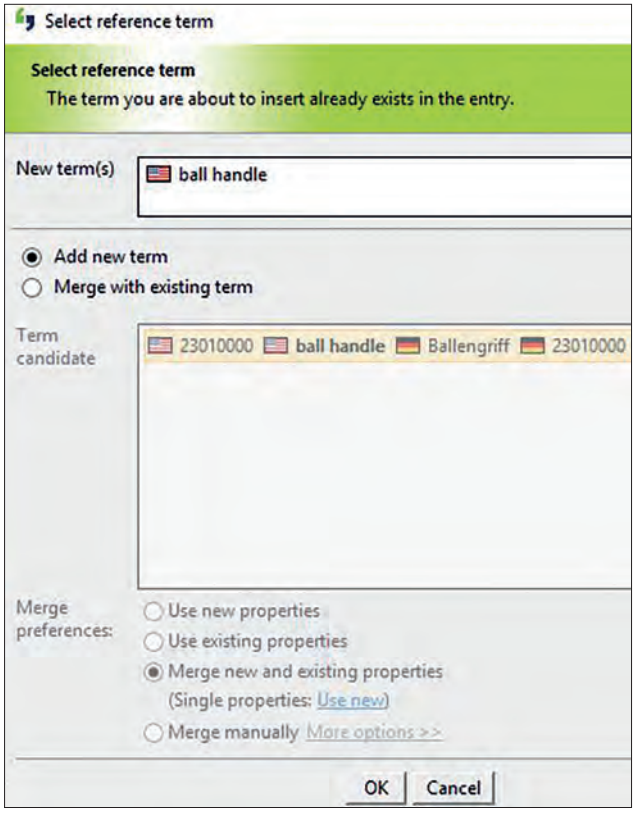


Fig. 4: Double-entry control at term entry in crossTerm (own figure)

Errors in content (= incorrect information) can occur in any data category. **Plausibility checks** ensure that a data record is corresponding content-wise with predefined criteria such

edition

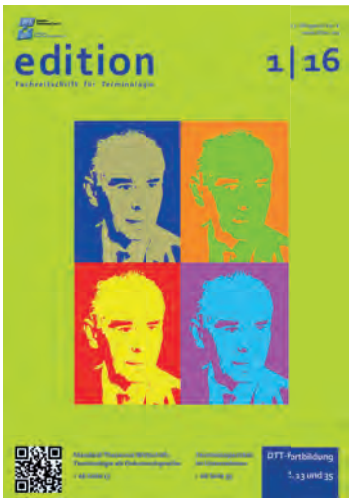
Fachzeitschrift für Terminologie

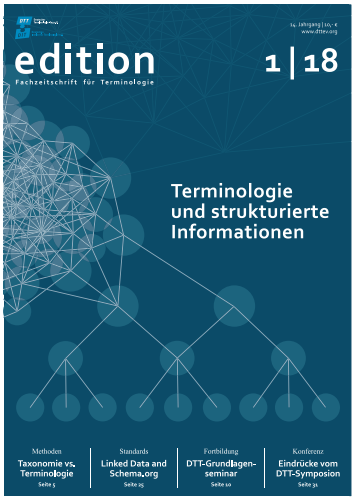
seit 2005

kostenloses Online-Archiv

Alle Ausgaben der edition finden Sie als PDF in unserem kostenlosen Online-Archiv. Perfekt zum Stöbern, Nachschlagen und Weitergeben.

www.dttev.org/edition





as the assignment of the correct source/target language. The **consistency check** determines whether terminological data conforms to specified criteria and whether interdependent terminological entries comply with each other. This applies to term formation and domain classification as well. Other consistency checks are reflecting the consistent use of data categories or the application of spelling rules in certain domains. The **linguistic check** is about the linguistic correctness (see ISO 23185 (2009)), e.g. the quality and correctness of the term. In addition, the definitions, the notes and the contexts are checked for editorial accuracy and whether domain-specific conventions have been applied.

It is advisable not to confuse data validation with term assessment criteria used to validate term candidates from the perspective of prescriptive terminology work (see Drewer/Schmitz 2017:80-90) such as e.g. brevity, univocality, motivation, conformity to laws and standards, usage, appropriateness, and internationality.

4.2 A brief insight into terminology tools

After having discussed the theory of helpful quality check methods, let us have a look into some tools and what kind of QA and QC support they offer.

One of the most important features is the so-called duplicate control. Most of the tools check for possible duplicate entries during the manual recording of a term and warn the user by an automated message when a possible duplicate is entered or saved. This feature is available with almost all commercial terminology management systems (e.g. MultiTerm, crossTerm, memoQ Translator Pro, QTerm, LookUp). Some commercial tools offer additional filters for duplicate entries. In MultiTerm it is called "Search for Duplicate Terms" and the search is carried out for homonyms (ambiguities) of the selected source language (see Fig. 5).

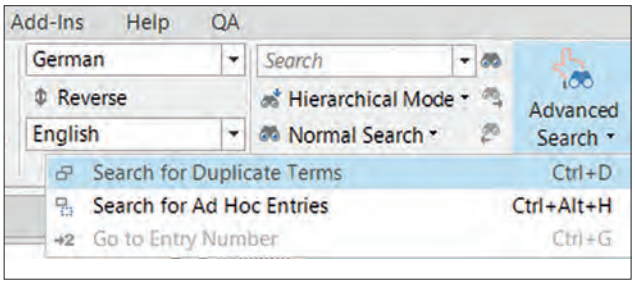


Fig. 5: Double-entry check with search filter for duplicate terms in MultiTerm (own figure)

LookUp offers also a separate ambiguity search, here the search can be done by language or simultaneously in all languages. In all cases, a list of ambiguous terms is displayed and the respective data entries can be processed directly from the list. LookUp offers another nice check feature as well. The tool checks where data in mandatory fields are missing (could be due to an import where data where missing) and it even lists outdated or wrong picklist values and marks the wrong picklist value in red (see Fig. 6).

A type of QA can also be warranted if you can customise predefined standard input templates with mandatory fields or predefined values in certain data categories to ensure consistent data entries. Some TMS offer mass operation features such as changing or adding picklist values to a large group of selected concepts or find and replace text in specific fields. With the crossTerm Mass Operations Wizard, one may add and remove picklist values of a selection of concepts and/or terms (see Fig. 7 on page 22). There is however no feature in crossTerm to search and replace text in certain text fields, whereas this can be done in MultiTerm with the optional Batch Edit menu. MultiTerm even offers a separate Batch Delete filter that deletes all units selected by use of a filter (see Fig. 8 on page 22). For MultiTerm, you also may use Excelling MultiTerm, a commercial tool that offers advanced termbase maintenance features to batch-repair termbase content or perform mass changes in a very comfortable and reliable way.

Unfortunately, no termbase system offers an integrated spell checker yet. So, to submit self-developed definitions and other textual information to a spell check, you need to use the Word spell checker for example and copy/paste the corrected text into your TMS, or export the termbase or part of its content later during a QC maintenance action and perform the spell check externally and then reimport the corrected data.

So, we have seen that some commercial TMS offer a few helpful QA and QC features. Nevertheless, when it comes to cleaning up, for example, a big database full of old data, or when different termbases need to be merged, it is more convenient to export data, consolidate them and then validate

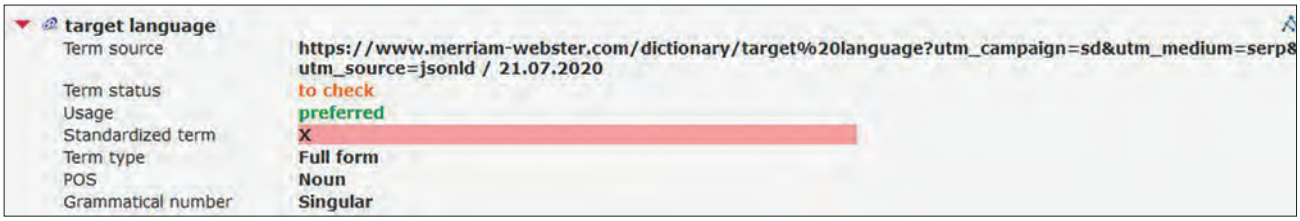


Fig. 6: Indication of a wrong data category value in LookUp 8 (own figure)

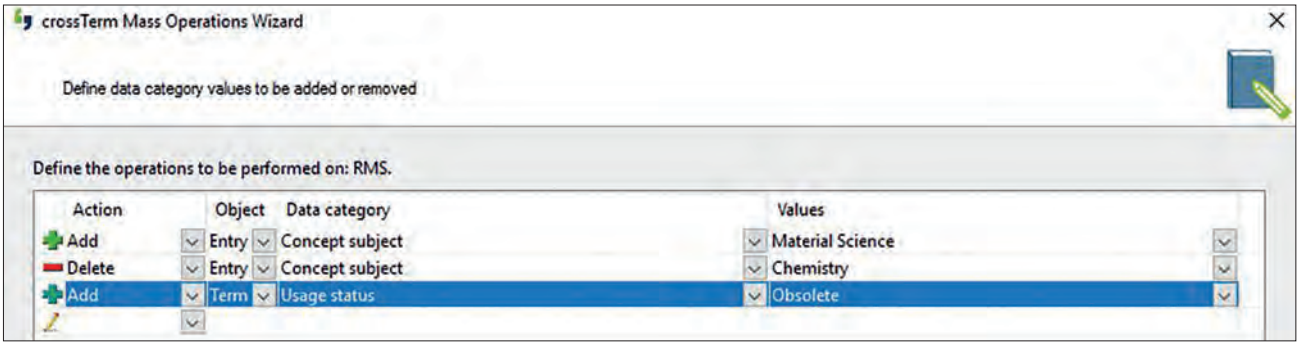


Fig. 7: Mass Operations Wizard in crossTerm (own figure)

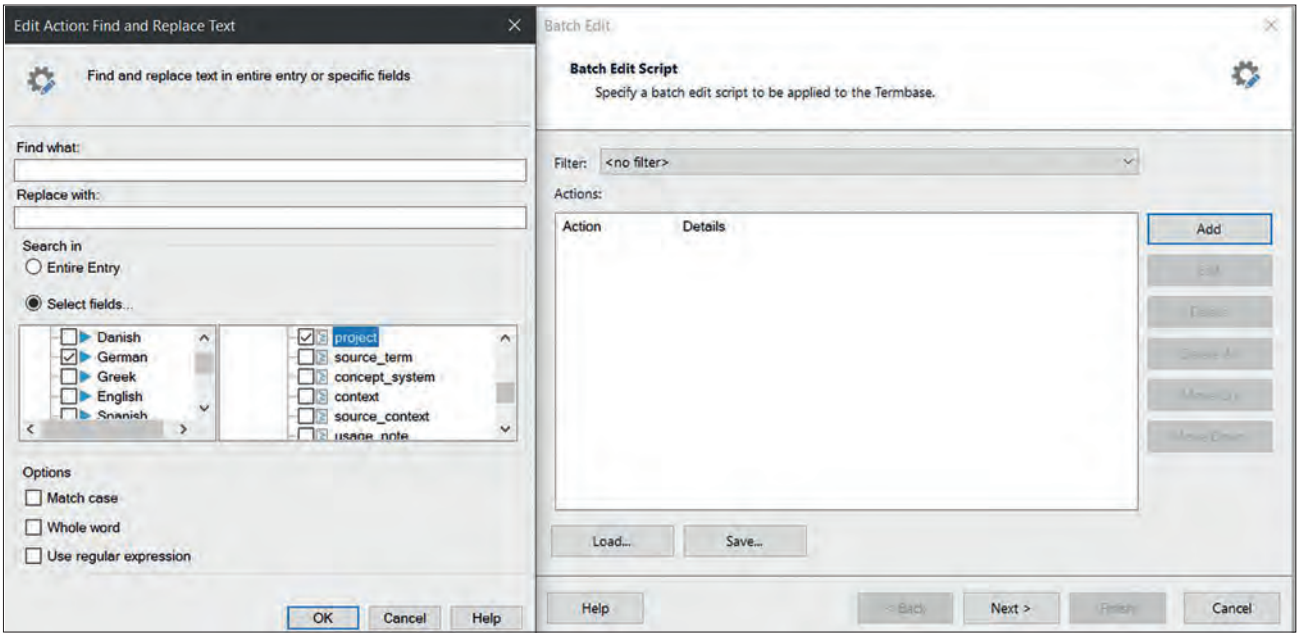


Fig. 8: Batch Edit Scripts to find and replace text in MultiTerm (own figure)

them in an external application, be it Microsoft Excel with the additional support of macros for certain regular check and clean-up options, or using other text editor tools.

5. Summary and Conclusions

Our aim was to present practical criteria to evaluate termbases grouped in four categories:

- environment,
- technical parameters,
- structure and content,
- usability and features.

It is sometimes difficult to distinguish clearly between aspects, as they can be closely interlinked; and user requirements of the various domains and aims place the emphasis on different editing principles. In our paper, we have also included some examples and a brief insight into terminology tools.

In case that no continuous database maintenance has taken place, it is helpful to perform an assessment of the termbase to find out which QC measures need to be taken to ensure that the termbase may fulfil its purpose to guarantee the required data quality in the future, too. The result of this assessment will help to schedule measures needed and coordinate the required human and financial resources.

This analysis cannot be considered as concluded, the termbases will necessarily change with newly emerging needs and the development of technology. With our paper we hope, we have been able to contribute to making a step forward on the way to an objective and practical evaluation of the quality of termbases.

References

COTSOES Conference of Translation Services of European States Working group "Terminology and Documentation" (Hrsg.) (2018): Empfehlungen für die Terminologiearbeit. (https://www.bk.admin.ch/bk/de/home/dokumentation/sprachen/publikationen-zur-terminologie.html) [accessed 19.09.2021].

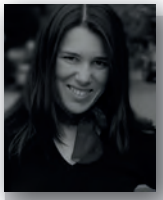
Drewer, Petra / Schmitz, Klaus-Dirk (2017): Terminologiemanagement Grundlagen - Methoden – Werkzeuge. Berlin: Springer-Verlag.

Fóris, Ágota / Rihmer, Zoltán (2007): A szótárak minősítési kritériumairól.

[On the Evaluation Criteria for Dictionaries]. In: Fordítástudomány 9(1), S. 109-113.
Frey, Dorina / Schmacht, Christine (2010): Aufbau eines Qualitätmanagements für die Terminologiearbeit. In: tekom-Jahrestagung und tc world conference 2010, Zusammenfassungen der Referate. Stuttgart: tcworld GmbH, S. 331-333.
Gaál, Péter (2012): Szempontrendszer online szótárak minősítéséhez. [Evaluation System for Online Dictionaries]. In: Magyar Terminológia 5(2), S. 225-250.
Hartmann, Reinhard R. K. (2001): Teaching and Researching Lexicography. Essex: Pearson.
ISO 704 (2009): Terminology work – Principles and methods. Genf: ISO.
ISO 1087 (2019): Terminology work and terminology science – Vocabulary. Genf: ISO.
ISO 12615 (2004): Bibliographic references and source identifiers for terminology work. Genf: ISO.
ISO 12616 (2002): Translation-oriented terminography. Genf: ISO.
ISO 12620 (1999): Computer applications in terminology – Data categories (withdrawn). Genf: ISO.
ISO 12620 (2009): Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources (withdrawn). Genf: ISO.
ISO 12620 (2019): Management of terminology resources – Data category specifications. Genf: ISO.
ISO 26162-1 (2019): Management of terminology resources – Terminology databases – Part 1: Design. Genf: ISO.
ISO 16642 (2017): Computer applications in terminology – Terminological markup framework. Genf: ISO.
ISO 23185 (2009): Assessment and benchmarking of terminological resources – General concepts, principles and requirements. Genf: ISO.
Kudashev, Igor (2013): Quality Assurance in Terminology Management: Recommendations from the TermFactory project. Helsinki: Unigrafia.
Ripfel, Martha (1989): Wörterbuchkritik. Eine empirische Analyse von Wörterbuchrezensionen. (Lexicographica. Series maior 29). Tübingen: Niemeyer-Verlag.
Sager, Juan C. (1990): A Practical Course in Terminology Processing. Amsterdam/Philadephia: John Benjamins Publishing Company, S. 141.
Schmitz, Klaus-Dirk (2020): Konzeption und Einrichtung von Terminologiedatenbanken. 12 Schritte zum Erfolg. In: edition, 01/2020, S. 11-17.
Schmitz, Klaus-Dirk (Coord.) (2020): Terminology Work Best Practices 2.0. Köln: Deutscher Terminologie-Tag e.V.
Steurs, Frieda / De Wachter, Ken / De Malsche, Evy (2015): Terminology tools. In: Kockaert, Henrik / Steurs, Frieda (Hrsg): Handbook of Terminology Management. Amsterdam/Philadelphia: John Benjamins, S. 228-248.
Tamás, Dóra Mária / Sermann, Eszter (2019): Evaluation System for Online Terminological Databases. In: Terminologija. Nr. 26/2019, S. 24-46. (http://lki.lt/wp-content/uploads/2020/03/Terminologija_26_maketas.pdf) [accessed 19.09.2021].
Tamás, Dóra Mária (2021): A WIPO Pearl szabadalmi terminológiai adatbázis bemutatása. [The Presentation of the WIPO Pearl Patent-related Terminological Database]. In: Fordítástudomány. 21. évfolyam 1. szám, S. 49-62. (https://ojs3.mtak.hu/index.php/fordtud/issue/view/607/335) [accessed 19.09.2021].
Tanke, Eberhard (2008): Die erste elektronische Terminologiedatenbank. In: Hennig, Jörg / Tjarks-Sobhani, Marita (Hrsg.): Terminologiearbeit für Technische Dokumentation (tekom Schriften zur Technischen Kommunikation, Band 12). Lübeck: Verlag Schmidt-Römhild, S. 32-36.
Warburton, Kara (2015): Managing terminology in commercial environments. In: Kockaert, Henrik / Steurs, Frieda (Hrsg): Handbook of Terminology Management. Amsterdam: John Benjamins, S. 381-382.
Warburton, Kara (Ed.) (2016): Terminology Starter Guide. (http://www.terminorgs.net/downloads/TerminOrgs_StarterGuide_V2.pdf) [accessed 15.09.2021].

Internet links:

bistro: http://bistrosearch.eurac.edu [accessed 18.09.2021]
Coreon: https://app.coreon.com [accessed 18.09.2021]
DatCatInfo: https://datcatinfo.termweb.se/termweb/app [accessed 18.09.2021]
EOHSTerm: https://eohsterm.org/sections/project/kb.php [accessed 18.09.2021]
EUR-Lex: https://eur-lex.europa.eu/ [accessed 18.09.2021]
Euskalterm: (http://www.euskadi.eus/web01-apeuskal/eu/q91EusTermWar/kontsultaJSP/q91aAction.do) [accessed 16.09.2021]
IATE: https://iate.europa.eu/home [accessed 18.09.2021]
LookUp: https://www.dog-gmbh.de/produkte/lookup-system-fuer-terminologiemanagement/ [accessed 18.09.2021]
Excelling MultiTerm: https://www.kaleidoscope.at/de/uebersetzungssoftware/experttools/excellling-multiterm/ [accessed 18.09.2021]
memoQ Translator pro: https://www.memoq.com/de/products/memoq-translator-pro [accessed 18.09.2021]
Multiterm: https://www.rws.com/translation/software/multiterm/ [accessed 18.09.2021]
Patentscope: https://patentscope.wipo.int/ [accessed 18.09.2021]
QTerm: https://portal.memoq.com/portal/de/professionelle-terminologie-verwaltung-qterm [accessed 18.09.2021]
SAPterm: http://www.sapterm.com/ [accessed 18.09.2021]
TERMDAT: www.termdat.ch [accessed 16.09.2021]
TERMIUM Plus: https://www.btb.termiumplus.gc.ca [accessed 16.09.2021]
WIPO Pearl: https://www.wipo.int/reference/en/wipopearl/ [accessed 17.09.2021]



Dóra Mária Tamás is a senior terminologist at the Hungarian Office for Translation and Attestation Ltd. (OFFI Ltd.), a researcher (PhD) and a teacher of terminology theory and practice in university courses. Her research focuses on translation-oriented terminology, harmonisation of national law as well as the analysis and classification of termbases.

Kontaktadresse
tamas.dora.maria@gmail.com

Beate Früh is a senior terminologist and owner of Büro b3 Terminologiemanagement, a company offering consultancy and coaching for terminology managemen as well as professional terminology services. She teaches terminology theory and practice at the FH Anhalt. In many customer projects she has encountered all types and qualities of termbases.

Kontaktadresse
frueh@buerob3.de
www.buerob3.de

Computergestützte Terminologieprüfung

Nicole Keller

Viele Sprachdienstleister, Unternehmen, aber auch freiberufliche Übersetzer erfassen inzwischen schon über Jahre hinweg systematisch ihre Benennungen in einem Terminologieverwaltungssystem und bauen sich damit eine wertvolle – meist mehrsprachige – Ressource auf, um nicht nur im Übersetzungsprozess, sondern auch direkt bei der Quelltexterstellung konsistenter arbeiten zu können. Diese Terminologiebestände können heutzutage in viele Prozesse zur Qualitätssicherung integriert werden, sodass sowohl Technische Redakteure – oder ganz allgemein „Quelltextersteller" – als auch Übersetzer, aber generell auch alle anderen Mitarbeiter eines Unternehmens konsistentere Texte entsprechend den Vorgaben verfassen können.

Zur Überprüfung der korrekten Verwendung von Benennungen in Quelltexten und Übersetzungen gibt es verschiedene Möglichkeiten, auf eine computergestützte Prüfroutine zurückzugreifen. Das beginnt bei integrierten, sprachübergreifenden Prüfungen in Translation-Management-Systemen, führt über eigenständige Tools zur Terminologieprüfung mit umfangreicheren, teils linguistischen Prüfroutinen bis hin zu rein linguistischen Systemen, die sich auf eine kleine Sprachauswahl konzentrieren.

Die Voraussetzungen für den sinnvollen Einsatz solcher Systeme und die Herausforderungen bei der korrekten computergestützten Terminologieerkennung sollen im Folgenden anhand verschiedener Beispiele und dreier exemplarischer Systeme erörtert werden.

Anforderungen an den terminologischen Eintrag

Für eine effektive, computergestützte Terminologieprüfung müssen einige grundlegende Anforderungen an die Strukturierung des terminologischen Eintrags erfüllt sein. So ist es zunächst essenziell, dass die Terminologiedatenbank bzw. der -bestand begriffsorientiert aufgebaut ist. Damit wird sichergestellt, dass Benennungen mit verschiedenen Bedeutungen nicht in einem einzelnen Eintrag gespeichert werden, sondern immer nur eine Bedeutung pro Eintrag existiert.

Darüber hinaus sollte es die Möglichkeit geben, den Verwendungsstatus einer Benennung hinterlegen zu können, sodass angegeben werden kann, ob eine Benennung bevorzugt, erlaubt oder verboten ist. Dies kann einerseits über eine frei konfigurierbare Datenkategorie mit dem Datenfeldtyp „Pickliste" erfolgen, die bspw. die Werte „bevorzugt", „erlaubt" und „verboten" enthält. Bei manchen Systemen ist die Verwendungsangabe auch bereits als feste Funktion integriert und bietet dann in der Regel die Optionen „bevorzugt" und „verboten" an. Im Falle von 1:1-Entsprechungen kann man im Übersetzungsprozess auf diese Zusatzangabe verzichten. Für die Quelltextprüfung sind jedoch verbotene Benennungen unverzichtbar, sodass nicht zulässige Synonyme gar nicht erst aus Versehen verwendet werden.

Außerdem ist es empfehlenswert, den Datenbestand nach Kunden zu strukturieren bzw. aufzuteilen, um zusammengehörige Datenbestände separat zu pflegen.

Im Falle eines mehrstufigen Terminologie-Workflows ist eine Statuskennzeichnung der Einträge bzw. Benennungen sinnvoll, damit auf den ersten Blick ersichtlich ist, in welchem Genehmigungsstadium sich diese befinden. So können beispielsweise noch nicht genehmigte Benennungen von der Prüfung ausgeschlossen werden.

Ansätze bei der Terminologieerkennung

Ein wichtiger Faktor bei der Terminologieerkennung ist die Erkennung von flektierten Formen im Satzgefüge. Diese korrekte Erkennung ist die Grundvoraussetzung für eine sinnvolle Prüfung. Ansonsten kommt es womöglich zu keiner Prüfung oder das System geht von einer ähnlich geschriebenen, für den Kontext aber falschen Benennung aus.

Grundsätzlich gibt es vier verschiedene Möglichkeiten, wie Benennungen im Satz erkannt werden, die im Folgenden kurz erläutert werden:

1. Exakte Erkennung: Hier wird die Benennung im Text nur erkannt, wenn die Schreibweise im Text exakt mit derjenigen in der Terminologiedatenbank übereinstimmt. Diese Option wird gerne bei Abkürzungen, Produkt- oder Eigennamen verwendet, da diese im Kontext nicht flektiert werden.
2. Stemming: Die derzeit häufigste Vorgehensweise bei sprachunabhängigen Prüfungen ist die Rückführung einer Benennung auf den „Wortstamm". Dies erfolgt aber in der Regel nicht auf der Basis von morphosyntaktischen Regeln, sondern dadurch, dass ein Wort um eine bestimmte Zeichenanzahl am Wortende gekürzt wird. So wird z. B. „Wanderung" als Benennung erkannt, wenn nur „wandern" als Eintrag in der Terminologiedatenbank gespeichert ist. Im Gegenzug werden

aber auch „sicher“ und „sichten“ in einen Topf geworfen, sodass mit diesem Ansatz auch falsche Ergebnisse produziert werden.
3. Fuzzy-Erkennung: Die unscharfe Worterkennung liefert überwiegend ähnliche Treffer wie das Stemming, erkennt aber zusätzlich leichte Unterschiede im Wort oder am Satzanfang, beispielsweise Tippfehler oder regional unterschiedliche Schreibweisen (BE: recognise vs. AE: recognize). Die Höhe des Ähnlichkeitsgrads für diese Erkennung kann in den meisten Systemen individuell eingestellt werden.
4. Linguistische Erkennung: Bei diesem Ansatz werden die Wörter und Wortbestandteile linguistisch analysiert und können somit auf den korrekten Wortstamm zurückgeführt werden. Auch unregelmäßige Flexionen stellen für diesen Ansatz kein Problem dar. Allerdings werden meist nur wenige Sprachen abgedeckt.

Terminologieprüfung und Herausforderungen in der Praxis

Für die Überprüfung auf korrekte Verwendung von Benennungen im Text gibt es nun in der Praxis einige Hürden zu nehmen, da die computergestützte Prüfung aufgrund der oben geschilderten Ansätze nicht immer optimal funktioniert. Folgende Fälle sind möglich:

1. Benennung im Quelltext korrekt erkannt (Prüfung möglich)
2. Benennung im Quelltext falsch erkannt (Verwechslung mit einem anderen Wort, Prüfung nicht möglich)
3. Benennung im Quelltext gar nicht erkannt (keine Prüfung)

Für den Übersetzungsprozess ergeben sich dieselben Probleme dann noch ein weiteres Mal für die Zielsprache.

Im Folgenden sollen einige Beispiele die Herausforderungen in der Praxis näher erläutern.

Flektierte Formen:
Ausgehend vom Deutschen stellen flektierte Formen für die Terminologieerkennung und -prüfung ein Problem dar, weil Benennungen dadurch häufig gar nicht oder falsch erkannt werden, wie z. B. schießen vs. schoss oder Buch vs. Bücher.

Komposita:
Im Falle von Komposita stellt die unterschiedliche Schreibweise je nach System eine gesonderte Herausforderung dar. So setzen manche Redakteure bereits sehr früh Bindestriche ein, wohingegen andere auch Komposita mit mehr als drei Wörtern zusammenschreiben (z. B. Telefonnummer vs. Telefon-Nummer). Dieser kleine Unterschied kann bereits zu einer Nichterkennung führen. Darüber hinaus können verbotene Benennungen Teil eines Kompositums sein und sollten auch als „verboten“ erkannt werden (siehe Beispiel im Absatz „Beispielsysteme“).

Abkürzungen:
Für die korrekte Erkennung von Abkürzungen ist es essenziell, dass ein System die Option „Groß-/Kleinschreibung beachten“ abdeckt. Ansonsten würde z. B. in einem deutschen Text bei der Abkürzung IST (Inverse Streutransformation) immer ein Treffer mit der Terminologiedatenbank erkannt werden, sobald im Satz …ist… vorkommt. Das wäre selbstverständlich kontraproduktiv.

Mehrwortbenennungen:
Benennungen, die aus mehreren Wörtern bestehen, können immer dann eine Herausforderung darstellen, wenn sie durch zusätzliche Informationen ergänzt werden, die zwischen den einzelnen Wörtern stehen, wie z. B. kombinierter Endpunkt vs. kombinierter, primärer Endpunkt, oder aber unerwartete Komposita gebildet werden, wie z. B. Schraube mit Vierkantloch vs. Vierkantlochschraube.

Situationsbedingte Verwendung:
Diese Herausforderung ist derzeit noch am schwersten zu lösen, da anhand von Metadaten festgelegt wird, wann eine konkrete Benennung zu verwenden ist. Das kann eine kundenspezifische Benennung sein, wie z. B. Blinker vs. Richtungsanzeiger, aber auch eine Benennung, die bis zu einem bestimmten Datum gültig war und danach durch eine neue abgelöst wurde. Aber auch zielgruppenspezifische Angaben (Fachsprache: Pankreas vs. Allgemeinsprache: Bauchspeicheldrüse) sind für ein System bisher nicht einwandfrei zuzuordnen. In diesem Fall muss immer noch der Mensch entscheiden, welche Benennung die richtige ist.

Beispielsysteme

Im Folgenden soll anhand der Terminologieprüfung in Trados Studio 2021 (RWS), checkTerm (Kaleidoscope) und Congree (Congree GmbH) gezeigt werden, wie sich die Terminologieprüfung stufenweise verbessern lässt.

Der verwendete Beispielsatz lautet: „Das Buch mit dem gelben Buchrücken steht neben den blauen Büchern im Bücherregal.“ In der angeschlossenen Terminologiedatenbank ist „Buch“ im Deutschen als „verboten“ markiert. Zu untersuchen ist, ob der unregelmäßige Plural und „Buch“ bzw. „Bücher“ in einem Kompositum als verbotener Bestandteil erkannt werden.

Trados Studio 2021:
Trados Studio ist ein Translation-Management-System, das für mehrsprachige Übersetzungsprojekte eingesetzt wird,

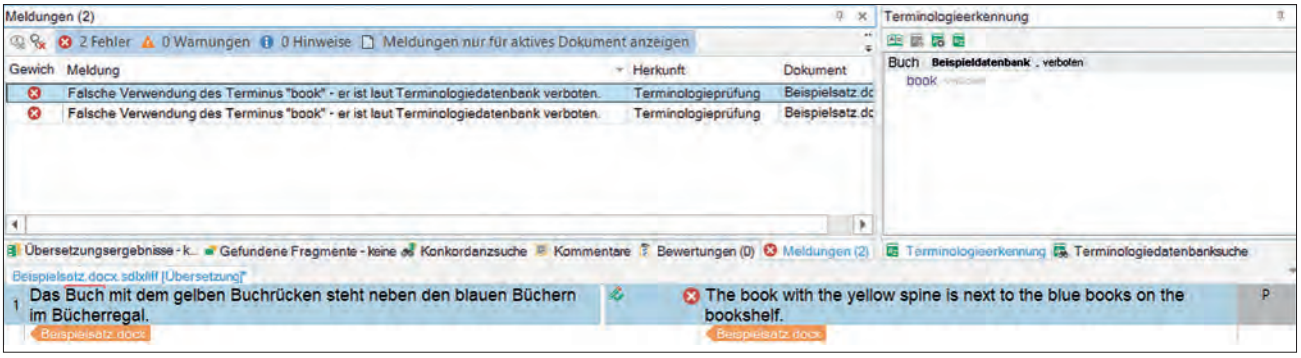


Abb. 1: Editor in Trados Studio 2021

eine eigene Terminologiedatenbank umfasst (MultiTerm) und im Übersetzungsprozess eine sprachunabhängige Terminologieprüfung einsetzt.

Abb. 1 zeigt den Editor von Trados Studio und die Übersetzung des oben genannten Beispielsatzes vom Deutschen ins Englische. Die Markierung im Ausgangssatz (rote Klammer über „Buch“) zeigt, dass das System die Benennung „Buch“ erkannt hat, aber alle weiteren Varianten von „Buch“ (Plural und Komposita) unentdeckt bleiben. Für den Zieltext wurde jeweils die Singular- und Pluralform von „book“ erkannt und als „Fehler“ markiert, das Kompositum „bookshelf“ wird übersehen.

checkTerm:

checkTerm ist ein eigenständiges Tool zur Terminologieprüfung, das auf die Daten in MultiTerm zugreift und entweder Text über die Zwischenablage oder direkt in verschiedenen Editoren über ein Plugin prüft. Aufgrund des morphologischen Ansatzes können auch leicht unterschiedliche Schreibweisen oder Zusammensetzungen korrekt erkannt und geprüft werden. checkTerm ist zunächst ein sprachübergreifendes Prüftool, verwendet aber für 20 Sprachen eine gesonderte Wortstammreduktion (u. a. für Deutsch), was wiederum einen größeren Mehrwert bietet.

In dem verwendeten Beispielsatz, bei dem in diesem Fall eine einsprachige Prüfung durchgeführt wurde, kann man erkennen, dass problemlos der unregelmäßige Plural von „Buch“ und auch das Kompositum „Buchrücken“ erkannt wurden und als „verboten“ auf den Eintrag mit der Benennung „Buch“ zurückgeführt wurden. Lediglich das Kompositum „Bücherregal“ wurde bei der Prüfung nicht erkannt. Im Vergleich zu Trados Studio ist das bereits ein erheblicher Fortschritt.

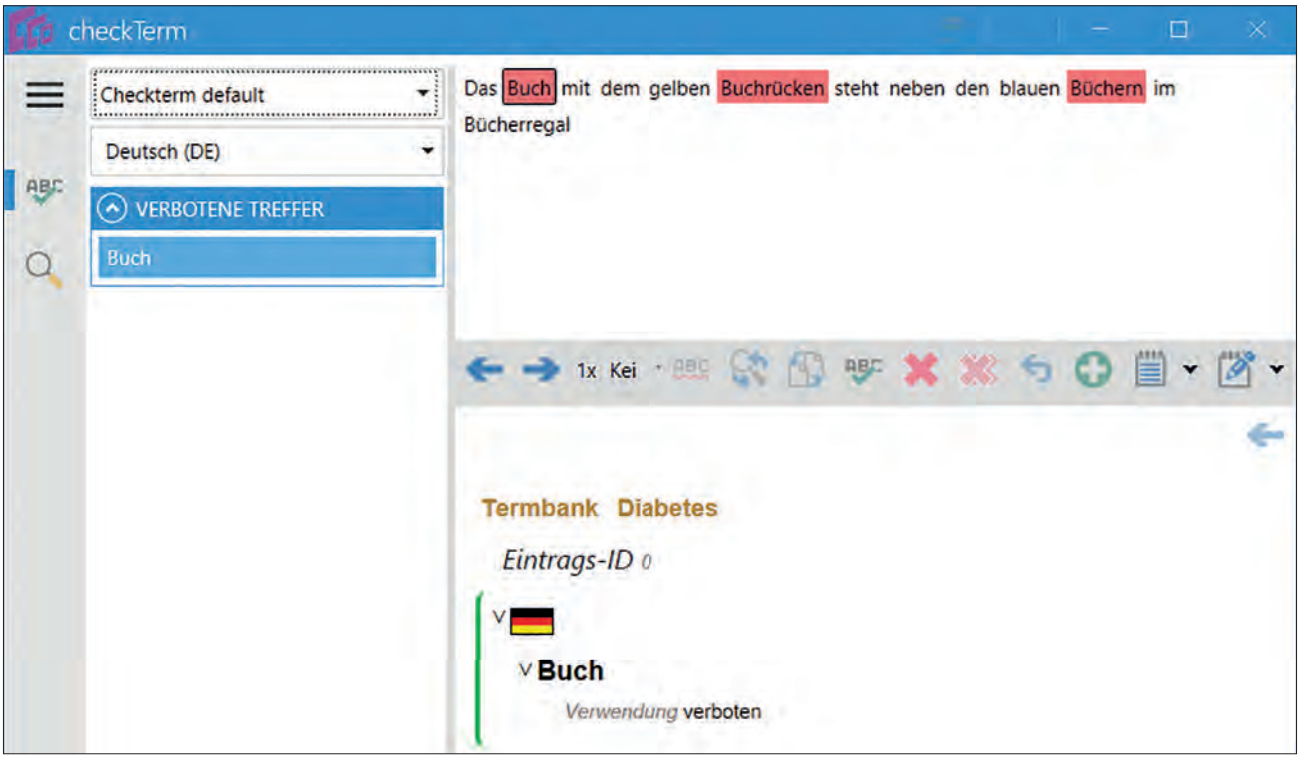


Abb. 2: checkTerm von Kaleidoscope

Abb. 3: Congree-Plugin in MS Word

Congree:

Der Congree Authoring Server wird zur Qualitätssicherung bei der Texterstellung eingesetzt und bietet neben einem Authoring Memory, das die Wiederverwendung von Sätzen ermöglicht, eine umfangreiche Sprachprüfung (Rechtschreibung und Grammatik, Stilprüfung, unternehmensspezifische Vorgaben) und eine Terminologieprüfung. Die Daten werden entweder in der internen Terminologiedatenbank gespeichert oder über ein Drittsystem in den Prüfprozess bei Congree eingebunden. Da Congree ein rein linguistisches Prüfsystem ist, werden auch nur eine Handvoll Sprachen derzeit abgedeckt. Die Prüfung erfolgt über verschiedene Plugins direkt in einem unterstützten Editor.

Abb. 3 zeigt, dass insgesamt 4 Terminologiefehler gefunden wurden, d. h., es wurden alle Fehler entdeckt – auch das Bücherregal. Ein weiterer Vorteil von Congree ist aber auch, dass es im Deutschen Mehrwortbenennungen auf Komposita zurückführen kann, wie z. B. Evaluierung der Daten vs. Datenevaluierung. Das bedeutet, dass das System darauf hinweist, dass es den Eintrag „Datenevaluierung" in der Datenbank gibt, falls Autoren „Evaluierung der Daten" im Text verwenden.

Fazit und Ausblick

Es wurde gezeigt, dass eine umfangreiche und sorgfältige Terminologiearbeit nicht unbedingt ein Garant für die korrekte Verwendung der Benennungen in Ausgangs- und Zieltexten ist. Es gibt sprachabhängig gesonderte Herausforderungen, die bei der Terminologieerkennung und damit auch bei der Terminologieprüfung zu bewältigen sind. Wie in vielen anderen Situationen kann die Maschine hier als Unterstützung miteinbezogen werden, den Menschen aber nicht ersetzen, sodass eine Humanüberprüfung oft unerlässlich ist.

Die verschiedenen Systeme haben allerdings gezeigt: Je umfangreicher die linguistische Komponente integriert wird, desto zuverlässiger fallen die Ergebnisse aus.

Dr. Nicole Keller ist Diplom-Übersetzerin und arbeitet als Dozentin am Institut für Übersetzen und Dolmetschen (IÜD) der Universität Heidelberg. Ihre Schwerpunkte liegen im Bereich Terminologiedatenbanken, CAT-Tools und Maschinelle Übersetzung. Sie arbeitet darüber hinaus als freiberufliche Übersetzerin im Bereich Medizin und gibt Schulungen für Übersetzer und Technische Redakteure.

Kontaktadresse
nicole.keller@iued.uni-heidelberg.de
www.iued.uni-heidelberg.de

Ergänzung zum Artikel „Visualisierung von Begriffsbeziehungen" in edition 1/21

Die Aussage, dass Begriffe bei Coreon nur in ein Repository aufgenommen werden können, wenn sie auch in ein Begriffssystem einsortiert werden können, soll hier noch einmal klar erläutert werden. Grundsätzlich können bei Coreon alle Begriffe aufgenommen werden, auch wenn sie kein Bestandteil eines bestehenden Begriffssystems sind. Für Problemgrößen wie Wortart, Adjektive usw. wird in der Praxis sehr häufig ein separater Knoten erstellt, um solche Einträge gebündelt zu erfassen. Das kann selbstverständlich auch für thematische Zusammenfassungen wie z. B. Prozessbeschreibungen erfolgen. Die Beschreibung im ursprünglichen Artikel war etwas unklar und sollte nicht suggerieren, dass in Coreon bestimmte Einträge gar nicht aufgenommen werden können.

Durchblick für Verbraucher

Norm für hochwertige Gebrauchsanleitungen veröffentlicht

Wie montiere ich das neue Regal, wie funktioniert die Kanalbelegung am Fernseher und was mache ich mit der kaputten Bohrmaschine? Gebrauchsanleitungen – treffender als „Nutzungsinformationen“ bezeichnet – liefern Verbrauchern Antworten auf diese und viele weitere Fragen. Das Deutsche Institut für Normung e.V. (DIN) hat in diesem Zusammenhang die DIN EN IEC/IEEE 82079-1 „Erstellung von Nutzungsinformationen (Gebrauchsanleitungen) für Produkte – Teil 1: Grundsätze und allgemeine Anforderungen“ veröffentlicht. Die Norm hilft Anbietern, qualitativ hochwertige Informationen zu ihren Produkten bereitzustellen. „Nutzungsinformationen sind als erstes zur Hand, wenn Verbraucher Produkte in Betrieb nehmen, Fragen zur Sicherheit haben oder Probleme lösen müssen“, sagt Dr. Gabriela Fleischer vom DIN-Verbraucherrat. „Untersuchungen zeigen jedoch, dass die Qualität hier stark schwankt. Die DIN EN IEC/IEEE 82079-1 soll dazu beitragen, dass Verbraucher zuverlässige Hinweise erhalten, um Produkte sicher, effizient und wirksam verwenden zu können.“

Breiter Geltungsbereich

Die DIN EN IEC/IEEE 82079-1 legt Anforderungen an Informationen für verschiedene Phasen im Lebenszyklus eines Produkts fest. Dazu zählen unter anderem Transport, Lagerung, Installation und Betrieb, aber auch Wartung, Reparatur sowie Recycling und Entsorgung. Fast jedes Produkt, das Verbraucher kaufen, enthält Nutzungsinformationen, deshalb deckt die Norm einen sehr breiten Bereich ab: Von Spielzeug über Haushaltsgeräte und Heimwerkerprodukte bis zu Autos gilt sie sowohl für elektrotechnische und nicht-elektrotechnische Produkte als auch für Software. Die allgemeinen Anforderungen zur Erstellung der Informationen berücksichtigen dabei verschiedene Zielgruppen – ob Verbraucher ohne spezielle Kenntnisse oder Fachleute. „Die Nutzungsinformation ist Teil des Produkts und somit auch Teil der Produktqualität“, erklärt Dr. Fleischer. „Wichtig für hochwertige und hilfreiche Informationen sind zum Beispiel ein logischer Aufbau, der gezielte und überlegte Einsatz von Warn- und Sicherheitshinweisen sowie verständliche Texte und Illustrationen.“

Begrifflichkeiten überarbeitet

Mit der DIN EN IEC/IEEE 82079-1 wurde die internationale Norm IEC/IEEE 82079-1 national übernommen. Experten aus acht Ländern hatten die Vorgängernorm IEC 82079-1 aus dem Jahr 2012 überarbeitet, dabei flossen auch Verbraucherinteressen stärker mit ein. In diesem Zuge wurde der Begriff der „Gebrauchsanleitung“ durch „Nutzungsinformation“ ersetzt, weil Gebrauchsanleitung zu eng gefasst war. Nutzungsinformationen enthalten nicht nur Angaben dazu, wie sich Produkte anwenden lassen, sondern auch referenzielle oder beschreibende Informationen. Sie umfassen unterschiedliche Informationsprodukte wie Anleitungen, Handbücher oder Servicepläne. Die Norm richtet sich unter anderem an Käufer und Anbieter von Produkten, Behörden und Sachverständige.

Die **DIN EN IEC/IEEE 82079-1** kann beim Beuth Verlag über www.beuth.de bezogen werden.

Weiterführende Informationen zur Norm sind auch auf der Seite der DKE in einer Artikelserie verfügbar: https://www.dke.de/de/arbeitsfelder/core-safety/nutzungsinformation-norm-als-grundlage

Über den Verbraucherrat

Der Verbraucherrat vertritt die Interessen der Endverbraucher in der nationalen, europäischen und internationalen Normung und Standardisierung. Er berät und unterstützt dabei die Lenkungs- und Arbeitsgremien von DIN. Das Bundesministerium der Justiz und für Verbraucherschutz (BMJV) fördert den Verbraucherrat auf Grund eines Beschlusses des Deutschen Bundestages.

Ausführliche Informationen unter: http://www.din.de/go/verbraucherrat

Das Gendern

Terminologisch betrachtet ein Missverständnis mit Skandalpotenzial

Lassen Sie mich terminologisches Denken in Kürze beschreiben: alles dreht sich um Dinge und deren Abgrenzung zu anderen Dingen. Und das im multilingualen Kontext. Klingt erstmal kompliziert – eigentlich ist es aber der Schlüssel, die Welt zu begreifen. So können Ihnen Terminologen die Eintönigkeit eines Kreuzworträtsels beweisen, sofern nicht eigentlich ein Wortkreuzrätsel gemeint ist und auch die Tatsache, dass „hinter Ihrem Rücken" eigentlich „vor Ihrer Brust" bedeutet. Und natürlich, warum Frauen nicht auf blauen Männerschemeln sitzen sollten. Lassen Sie mich das heute einmal am Beispiel des Genderns vorführen und ein großes Missverständnis aufklären.

Punkt 1: Die Sache mit den Homonymen: Warum ein Arzt nicht ein Arzt und ein Lokführer nicht ein Lokführer ist.

Wir alle kennen noch das gute alte Teekesselchen-Spiel: „auf das eine Teekesselchen bring ich mein Erspartes und auf dem anderen Teekesselchen sitzt Oma im Park". Na klar, die Bank! Nun sind aber nicht alle Homonyme derart leicht zu identifizieren, denn die deutsche Sprache birgt zahlreiche versteckte Mehrdeutigkeiten. Ich würde bezweifeln, dass jemals folgende Teekesselchen-Aufgabe formuliert wurde, obwohl es denkbar und richtig wäre: „Das eine Teekesselchen ist eine Person mit medizinischem Sachverstand und der Verantwortung, diesen zu meinem Wohl einzusetzen, sofern ich erkranke. Und das andere Teekesselchen ist eine männliche Person mit medizinischem Sachverstand und der Verantwortung, diesen zu meinem Wohl einzusetzen, sofern ich erkranke." Lesen Sie ruhig noch einmal drüber. Und dann lesen Sie gern noch einmal drüber. Solange, bis Sie mir zustimmen:

Arzt und Arzt sind zwei unterschiedliche Begriffe. Und genauso verhält es sich mit Lokführer und Lokführer und mit Teilnehmer und Teilnehmer und mit Kellner und Kellner und mit Steuerberater und Steuerberater und... Und die jeweiligen Homonyme unterscheiden sich dadurch, dass der eine Begriff die geschlechtsneutrale Berufsbezeichnung präsentiert und der andere Begriff Wert auf die Nennung und Betonung des Geschlechtes legt. Quasi: „eine Person mit medizinischer Kompetenz" vs. „eine Person und medizinischer Kompetenz UND männlichen Geschlechtsmerkmalen". Da sorgt die deutsche Sprache durch den Hang zum generischen Maskulinum also für Missverständnisse.

Punkt 2: Die Sache mit den Attributen: Warum Eigenschaften besser nicht Teil der Benennung sein sollten.

Was für eine Rolle spielt aber nun das Geschlecht? Nun, Sie wissen selbst, dass die Antwort durchaus kontextabhängig ist. Im Rahmen von Berufsbezeichnungen ist sie (ja, auch hier gibt es Ausnahmen) in der Regel eindeutig: Das Geschlecht ist von keinerlei Bedeutung in Bezug auf die Ausübung eines Berufes. Denn ein Penis sollte genauso viel Einfluss auf eine Frisur haben, wie eine Vagina auf die Besohlung meiner Schuhe: keinen. Terminologisch betrachtet ist das Geschlecht damit eine Eigenschaft, kein Merkmal. Lassen Sie mich das an einem anderen Beispiel veranschaulichen.

Zentrale Frage aller Terminologie: Was macht die Sache zu dem, was sie ist? Was macht einen Stuhl zum Stuhl, einen Hocker zum Hocker oder einen Schemel zum Schemel? Ein Stuhl definiert sich im Allgemeinen als Sitzgelegenheit mit vier Beinen und Rückenlehne. Nimmt man dem Stuhl die Rückenlehne, wird er zum Hocker. Nimmt man ihm zusätzlich ein Bein und sägt die anderen ein bisschen ab, so wandelt er sich zum Schemel. Striche man den Schemel nun mit blauer Farbe, dann... bliebe er ein Schemel.

Es gibt also wesenseigene Attribute, die eine Sache zu einer Sache machen, weil sie sich durch diese wesentlich von anderen Dingen abgrenzen: die Merkmale. Und es gibt Attribute, die das Wesen einer Sache nicht tangieren: die Eigenschaften.

Da das Geschlecht das Wesen eines Berufes überhaupt nicht beeinflusst, muss es sich also um eine Eigenschaft handeln. Und Eigenschaften sind in der Regel nicht Bestandteil der Benennung. Oder kennen Sie einen Männerstuhl? Oder einen Frauenschemel? Oder einen blauen Männerhocker? Wer auf dem Stuhl Platz nimmt, ist für das Ding und seine Benennung genauso wenig ausschlaggebend wie die Farbe. In Bezug auf Berufe oder Rollen sollte also auch das Geschlecht nicht Bestandteil der Benennung

sein. Denn neben dem Geschlecht gibt es zahlreiche andere Eigenschaften: Haarfarbe, Schuhgröße, Augenfarbe, Lieblingsfilm, etc. Und alle haben keinerlei Auswirkung auf die Berufsausübung. Wenn man also zwischen Arzt und Ärztin unterscheidet, dann könnte man auch zwischen gelbbesocktem Arzt mit Halbglatze und Faible für Rockmusik und der muslimischen Ärztin mit Vorliebe für Yoga und Heavy Metal unterscheiden. Sie sehen – es ergibt einfach keinen Sinn.

Punkt 3: Der (sprach-)logische Lösungsansatz. Und das Wittern der Gefahr.

Wenn wir also nun erkannt haben, dass das Geschlecht eine Eigenschaft ist, dann müssen wir auch erkennen, dass Diskriminierung nicht aufgehoben werden kann, indem man alle Geschlechter in einer Berufsbezeichnung mitnennt. Und wenn wir eventuelle Missverständnisse vermeiden wollen, die die deutsche Sprache durch Verwendung des generischen Maskulinums mit sich bringt, dann sollten wir nicht in die Falle tappen und künstliche, eigenschaftsbasierte Benennungen kreieren, damit „alle mitgenannt“ sind. Vielmehr wäre das Gegenteil die Lösung: Wir sollten Benennungen schaffen, die merkmalbasiert sind. Also: genderneutral.

Statt Lokführer und Lokführerinnen oder statt Lokführer:Innen oder statt Lokführer*innen sollten Berufe durch eine geschlechtsneutrale Form repräsentiert werden. Es gibt diesbezüglich Vorschläge zuhauf, die zunächst noch etwas gewöhnungsbedürftig klingen, durch ihre (sprach-)logische Unantastbarkeit allerdings auf absehbare Zeit langfristig Eingang in unsere Sprache finden werden: z. B. das Arzty, das Lokführy, etc.

Als Terminologe bin ich überzeugt, dass nur eine geschlechtsneutrale Form dauerhaft dafür sorgen wird, Diskriminierung zu vermeiden. Zu groß ist die Gefahr, dass die geschlechtsspezifische Trennung von Arzt und Ärztin dafür sorgen könnte, dass manche Menschen ausschließlich von Ärzten operiert werden wollen...

Tom Winter
winter@dttev.org

Ist Gott weiblich?

Überarbeitung der Bibel in gerechter Sprache

Das Thema „Gendern“ zieht auch im kirchlichen Umfeld seine Kreise, z. B. mit der Bibel in gerechter Sprache. Sie ist eine Bibelübersetzung mit dem Ziel, die biblischen Schriften aus den Ursprungssprachen so in die deutsche Gegenwartssprache zu übertragen, dass sie auch der Bedeutung der Frauen in der Bibel gerecht wird und gegenüber dem Judentum sensibel ist. Sie wurde 2006 auf der Frankfurter Buchmesse vorgestellt und stieß damals wie heute sowohl auf Zustimmung als auch auf Kritik.

15 Jahre nach der Erstveröffentlichung soll nun die Arbeit an einer Neufassung beginnen. Dabei sollten auch die Debatten um Geschlechtergerechtigkeit und Postkolonialismus der vergangenen Jahre berücksichtigt werden. Unter anderem soll in diesem Zusammenhang auch bei der Bibelübersetzung die Frage nach dem „Gendersternchen“ diskutiert werden. Wir dürfen gespannt sein.

Glossar

Für Terminologen interessant dürfte das Glossar auf der Website der Bibel in gerechter Sprache sein. In diesem Glossar sind hebräische und griechische Wörter mit gleichen oder ähnlichen Bedeutungsfeldern oft in einem Artikel zusammengefasst. Das lässt nicht selten Unterschiede zwischen den jeweiligen Sprachwelten deutlich werden, zeigt aber zugleich den literarischen, sozialen und den grundlegenden theologischen Zusammenhang der verschiedenen Teile der Bibel. In einigen Fällen sind mehrere Worte und Wortfelder unter deutsche Sammelbegriffe gefasst (z. B. Opfer, Gerechtigkeit). Auf diese Weise können begriffliche Zusammenhänge und Differenzen kompakt erfasst werden. Zwei (willkürlich gewählte) Beispiele:

Gott, Gottesnamen, Gottesbezeichnungen
Ein aufmerksames Umgehen mit dem Gottesnamen gehört zu den zentralen Anliegen der Bibel in gerechter Sprache. Das bezieht sich vor allem auf den Eigennamen Gottes, der im Ausgangstext mit den Konsonanten j-h-w-h geschrieben wurde. Um dessen Heiligkeit zu wahren, wurde er seit biblischer Zeit nicht ausgesprochen, sondern durch ein Ersatzwort wiedergegeben. Die Aufmerksamkeit für den Gottesnamen drückt sich in der Bibel in gerechter Sprache in den Übersetzungen der unterschiedlichen Namen aus und in einer besonderen graphischen Wiedergabe des Eigennamens Gottes in Verbindung mit der Kopfzeile über jeder linken Seite des Bibeltextes.

Bote, Botschaft, Engel, Engel(s)gestalt – malach (hebr.); angelos (griech.)
Die Bibel in gerechter Sprache zielt mit den gewählten Formen der Wiedergabe darauf, dieser Vielfalt biblischen Redens von malach / angelos gerecht zu werden. Dabei wird klar, dass jede Übersetzungsentscheidung (neue) Fragen aufwirft: Wie kann deutlich werden, dass das Geschlecht von malachim / angeloi nicht feststeht, wenn malach / angelos mit Bote wiedergegeben wird? Haben malachim / angeloi, sofern sie nicht irdischer Herkunft sind, überhaupt ein Geschlecht? Und verschärft sich das andere Problem, nämlich das der Sonderstellung von Botenwesen nicht irdischer Herkunft, womöglich sogar, wenn malach / angelos mit Engelsgestalt wiedergegeben wird, um in der deutschen Sprache ein feminines Wort zu verwenden?

Wer weiter stöbern möchte, findet das Glossar hier: https://www.bibel-in-gerechter-sprache.de/die-bibel/glossar/

Quellen:

[1] Wikipedia: https://de.wikipedia.org/wiki/Bibel_in_gerechter_Sprache [Zugriff am 02.11.2021]
[2] Deutschlandfunk: https://www.deutschlandfunk.de/genderdebatte-die-bibel-in-gerechter-sprache-wird.2849.de.html?drn:news_id=1311121 [Zugriff am 02.11.2021]
[3] Bibel in gerechter Sprache: https://www.bibel-in-gerechter-sprache.de/ [Zugriff am 02.11.2021]

Zum Thema Gendern gibt es auch noch diese interessante Studie, die über den angegebenen Link erreichbar ist:
Kricheli-Katz, Tamar, Tali Regev (2021): „The effect of language on performance: do gendered languages fail women in maths?“ In: *npj Science of Learning*, Ausgabe 6, 9 (2021). https://doi.org/10.1038/s41539-021-00087-7 [Zugriff am 02.11.2021]

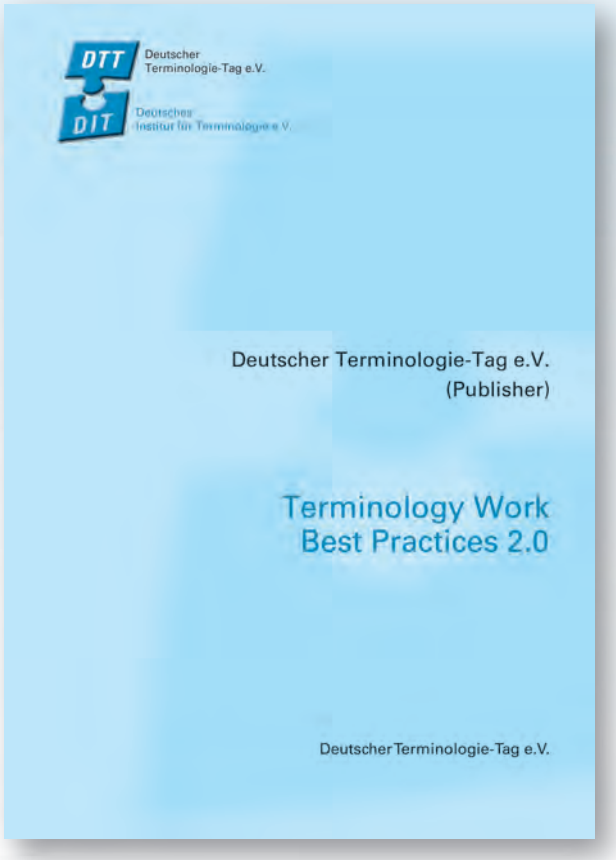
Deutscher
Terminologie-Tag e.V.

Deutsches
Institut für Terminologie e.V.

Terminology Work Best Practices 2.0

Das bewährte Praxishandbuch für Terminologiearbeit und Terminologiemanagement mit dem Know-how zahlreicher Experten aus Industrie und Wissenschaft ist jetzt **erstmals in englischer Sprache** erhältlich.

kompakt | praxisnah | auf den Punkt



Geballtes Terminologiewissen für Unternehmen und Freiberufler

1. Argumentationshilfen
2. Grundsätze und Methoden
3. Benennungen
4. Werkzeuge und Technologien
5. Projekt- und Prozessmanagement
6. Berufsprofile, Anforderungen, Ausbildungsinhalte
7. Urheberrecht an Terminologie
8. Wirtschaftlichkeit

Erhältlich in Deutsch oder Englisch
auf www.dttev.org

Wasser oder Wässer – ein Fall für Terminologen?

Die unscheinbare Datenkategorie Numerus wird in Terminologiedatenbanken und beim Erfassen von Benennungen häufig recht stiefmütterlich behandelt. Es gibt zwar Sonderfälle, zum einen Unternehmen, die aus ihrer Terminologiedatenbank ein linguistisches Terminologieprüftool wie Congree oder Acrolinx füttern und dann oft diese Daten als „single source of truth" pflegen und ins Prüftool geben, damit Daten nicht unabhängig voneinander an zwei Stellen gepflegt werden. Und zum anderen Terminologen, die unregelmäßige Pluralformen in ihre Terminologiedatenbank einpflegen und diese dann entsprechend kennzeichnen. Aber meistens zieht das klassische Argument, dass man als Muttersprachler ja den Numerus ohnehin kennt, die kanonische Form meist sowieso der Singular ist und das Befüllen der Datenkategorie nur einen unnötigen zusätzlichen Zeitaufwand bedeutet.

Stille Wasser sind tief

Bei der Datenkategorie Numerus handelt es sich jedoch um eine weitaus komplexere Datenkategorie, als man auf den ersten Blick meinen könnte. Definiert wird der Numerus als die grammatische Unterscheidung, die die Anzahl der Objekte angibt, auf die sich eine Benennung bezieht [1]. In den meisten Terminologiedatenbanken sind die einzigen Picklistenwerte, die hierbei zur Auswahl stehen „Singular" oder „Plural", manchmal gibt es vielleicht auch noch „andere". Das muss doch reichen, könnte man meinen. Dem ist jedoch nicht unbedingt so.

Die kanonische Form, in der im Deutschen eine Benennung in einer Terminologiedatenbank erfasst wird, ist der Nominativ Singular. Das funktioniert meist gut, bis man auf eine Benennung wie beispielsweise „Kosten" trifft. Es handelt sich augenscheinlich um eine Pluralform und wird deshalb als solche ausgezeichnet. Jedoch existiert das Wort „Kosten" nicht im Singular, was es zu einem Pluraletantum macht. Analog gibt es Singularetantums, also Wörter die ausschließlich im Singular vorkommen wie beispielsweise „Vieh". Solche Fälle sollten ohne Frage auch in einer Terminologiedatenbank gekennzeichnet werden, besonders wenn diese auch von Personen benutzt wird, deren Muttersprache nicht Deutsch ist.

Ein Glas, zwei Gläser – alles glasklar, oder?

Eine weitere Ausnahme sind Stoffnamen. Diese zeigen an, dass es sich um eine einheitliche, nicht unterteilbare Entität handelt, wie dies unter anderem bei „Salz", „Holz" oder „Glas" der Fall ist. Spricht man zum Beispiel vom Material „Glas", so wird das Wort stets im Singular verwendet, denn es ist nicht zählbar, außer es geht beispielsweise um die chemische Zusammensetzung von verschiedenen „Gläsern", also „Glasarten". Dabei muss jedoch beachtet werden, dass dieselbe Benennung ggf. auch noch als zählbarer Begriff „Glas" vorkommen kann, der synonym zu bzw. als Abkürzung von „Trinkglas" verwendet wird. Da er zählbar ist, wird selbstverständlich der Plural „Gläser" verwendet. Wird „Gläser" jedoch im Sinne von „Glasarten" verwendet, so handelt es sich um einen Fachplural, genau wie auch der Fachplural „Wässer" spezielle Wasserarten, z. B. Mineralwässer, beschreibt und der Plural „Wasser" hingegen Wassermassen oder Gewässer meint [2]. Solche Fälle sollten in der Terminologiedatenbank vermerkt werden. Umgekehrt gibt es auch Fachsingulare, die meist sehr ungewohnt klingen, aber dennoch existieren. Beispiele hierfür sind „das Elter" oder „das Geschwister".

Auszeichnung in der Terminologiedatenbank – Platzsparen vs. Normenkonformität

Als Terminologe steht man nun vor dem Problem, wie man beim Eintragen eines Pluraletantums, Singularetantums, Stoffnamens, Fachsingulars oder Fachplurals am besten vorgeht: Eine Möglichkeit wäre natürlich, zusätzlich zu den bestehenden Picklistenwerten „Singular", „Plural" (und ggf. „andere") noch weitere Werte hinzuzufügen. Daraus könnte man dann „Singularetantum", „Pluraletantum" oder „Stoffname" wählen und die Benennung wäre korrekt ausgezeichnet. Jedoch können nicht in allen Terminologieverwaltungssystemen bei vorgegebenen Datenkategorien wie „Numerus" eigene Picklistenwerte hinzugefügt werden. Und selbst wenn die Möglichkeit besteht, so geht es eventuell auf Kosten der Übersichtlichkeit, drei zusätzliche Picklistenwerte für doch eher seltene Fälle anzulegen.

Eine weitere Möglichkeit besteht darin, nur Singular und Plural als Picklistenwerte zur Auswahl zu haben, und sich ein Anmerkungsfeld zunutze zu machen. So könnte man zum Beispiel ein extra Textfeld „Anmerkung zum Numerus" anlegen, in das man dann Kommentare wie „Singularetantum", „Pluraletantum", „Stoffname" oder z. B. auch Hinweise auf eine irreguläre oder fachsprachliche Pluralbildung eintragen kann. Natürlich geht auch die Option, eine weitere Datenkategorie zu erstellen, zulasten der

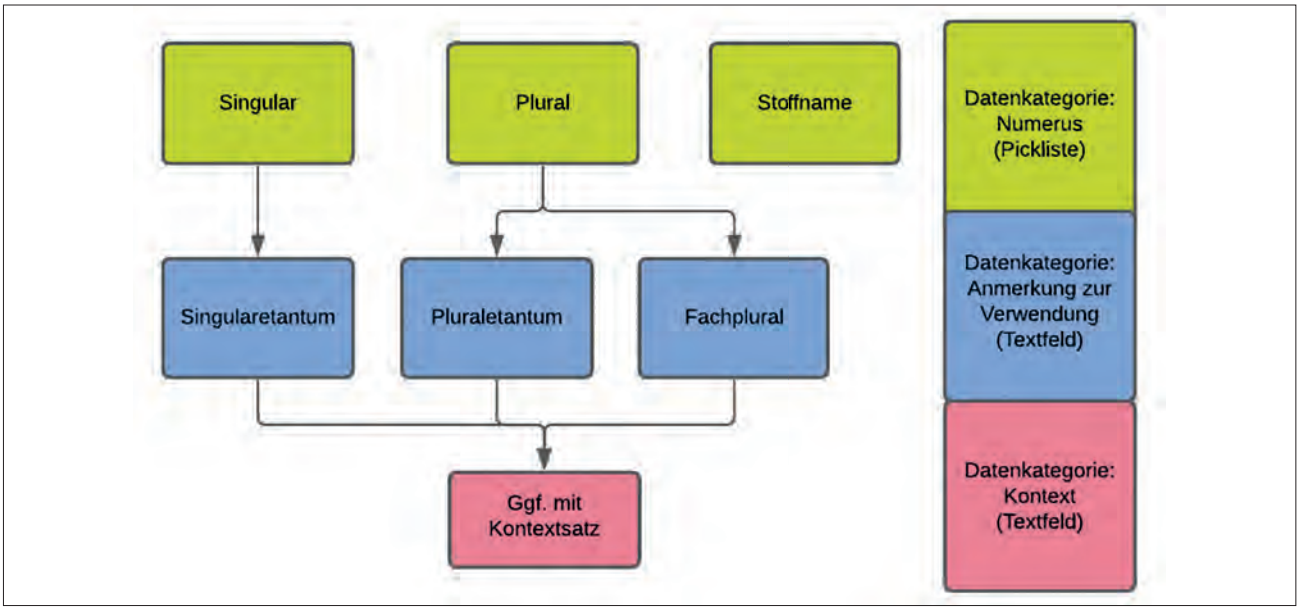


Abb. 1: Vorschlag für Informationen und betroffene Datenkategorien beim Auszeichnen des Numerus im Deutschen (eigene Darstellung)

Übersichtlichkeit. Jedoch ist auf diese Weise immerhin das Kriterium der Elementarität gewahrt, was bei der dritten Option, nämlich anstatt einer eigenen Datenkategorie „Anmerkung zum Numerus" eine bereits bestehende Datenkategorie wie „Anmerkung zur Verwendung" für Kommentare zum Numerus zu verwenden, nicht unbedingt der Fall ist. Welche Datenkategorien bei dieser Variante betroffen sind und mit welchen Informationen sie gefüllt werden, ist in Abbildung 1 zusammengefasst.

Diese Option funktioniert gut, solange man nicht noch eine weitere Anmerkung zur Verwendung, die nichts mit dem Numerus zu tun hat, wie z. B. „nur in Marketingtexten verwenden" eintragen möchte. Wenn sich nicht an Elementarität gehalten wird, lässt sich deutlich schlechter filtern.

Numerus ist nicht gleich Numerus – Was tun in bilingualen Termbanken?

Auch im Englischen kommt der Numerus nicht ganz so einfach daher. Mass nouns (in Großbritannien spricht man von collective nouns), wie z. B. „evidence" oder „luggage", bezeichnen etwas Unzählbares, das abstrakt oder unbestimmbar ist [3]. Diese Information ist zweifelsohne wertvoll in einer Terminologiedatenbank. Zusätzlich können jene mass nouns als only singular oder only plural form auftreten, wie das bei „courage" oder „manners" der Fall ist, was in Terminologieeinträgen ebenso vermerkt werden sollte. Die vermeintlich „normalen" Singular- und Pluralformen können in der englischen Sprache unter Umständen auch Besonderheiten aufweisen. Bei den Substantiven „police"

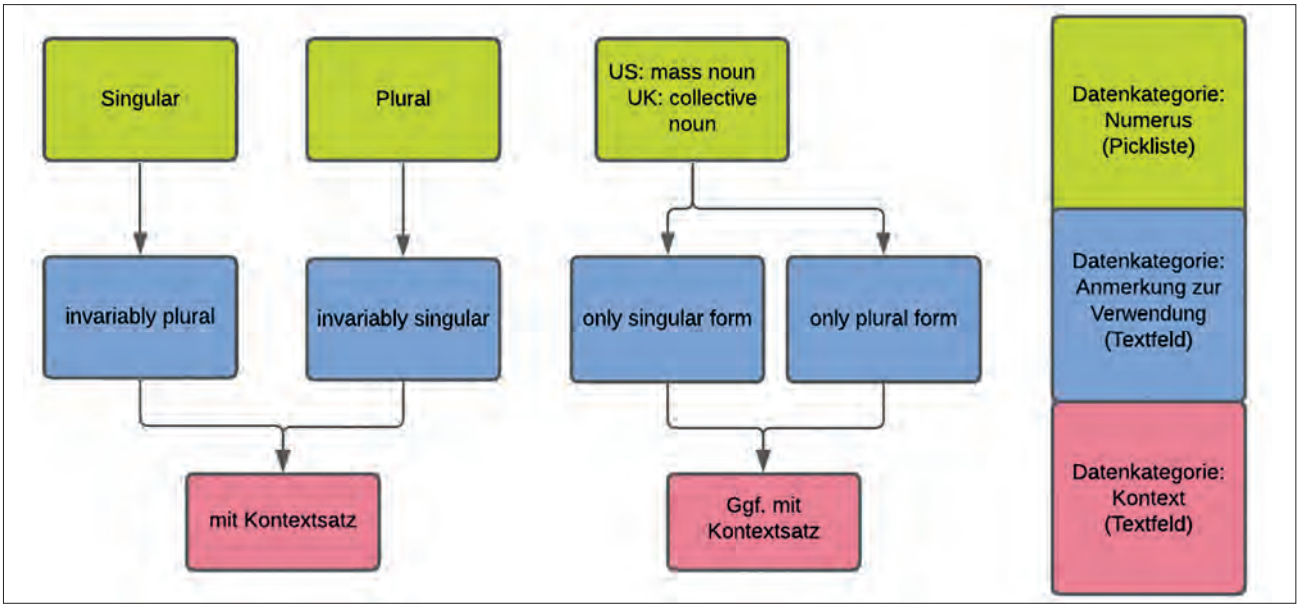


Abb. 2: Vorschlag für Informationen und betroffene Datenkategorien beim Auszeichnen des Numerus im Englischen (eigene Darstellung)

oder „news“ hat man es mit invariably plural bzw. invariably singular zu tun, d. h., dass die Substantive in einer Form erscheinen (also mit oder ohne Plural-s am Wortende), aber im Satz gegensätzlich verwendet werden: „The police were helpful.“ vs. „The news is good.“.

In bilingualen Terminologiedatenbanken sind besondere Numerusformen der beiden Sprachen nicht deckungsgleich. Die englische Benennung „instructions“ muss als mass noun und only plural form ausgezeichnet werden, während die deutsche „Anleitung“ mit der Information Singular ausreichend gekennzeichnet ist. Dies stellt sowohl bei der Datenbankkonzeption als auch bei neuen Einträgen eine große Herausforderung dar: Soll für jede Information eine eigene Datenkategorie angelegt werden? Wie oft braucht man diese speziellen Informationen überhaupt? Um in bilingualen Terminologiedatenbanken die Übersichtlichkeit zu wahren, bietet es sich an, bestehende Datenkategorien für derart spezielle Informationen zu verwenden. Wie sich das für das Englische darstellen könnte, ist in Abbildung 2 (Seite 34) zu sehen.

Fest steht, dass man in zweisprachigen Terminologiedatenbanken Kompromisse eingehen muss. Deutsche Stoffnamen sind nicht identisch mit den englischen mass nouns, dennoch bietet sich ein geteilter Picklistenwert an. Gleiches gilt für die Datenkategorie Anmerkung zur Verwendung, da diese Lösung bei unterschiedlichen Numerusausprägungen universeller ist. Platz sparen hinsichtlich der Numerusauszeichnung scheint also vor allem in bilingualen Terminologiedatenbanken sinnvoll.

Zusammenfassung

Egal, ob man sich nun für volle Normenkonformität oder die platzsparende Variante entscheidet, wichtig ist, dass dem Numerus in der Terminologiearbeit der nötige Platz eingeräumt wird. Dabei können beispielsweise auch Kontextsätze eine Rolle spielen, um die speziellen Numerusformen zu veranschaulichen. Terminologen müssen sich in jedem Fall mit dem Numerus ihrer Sprache(n) auseinandersetzen und eine konsistente Lösungsstrategie bei dessen Erfassung erarbeiten. Denn nur so sind sie mit allen Wassern gewaschen – oder Wässern?

Quellenverzeichnis

[1] DatCatInfo http://datcatinfo.termweb.se/datcat/DC-251 [15.09.2021]
[2] Duden Sprachwissen https://www.duden.de/sprachwissen/sprachratgeber/Die-Wasser-oder-die-W%C3%A4sser-Verschiedene-Pluralformen [15.09.2021]
[3] Chicago Manual of Style https://www.chicagomanualofstyle.org/book/ed17/part2/ch05/psec007.html [15.09.2021]

Franziska Fischer und Pascal Müller
fischer@buerob3.de
pascal.mueller26@gmail.com

20. Internationaler EURALEX-Kongress 2022 in Mannheim

Unter dem Motto „Wörterbücher und Gesellschaft“ (Dictionaries and Society) wendet sich die Konferenz an Expertinnen und Experten aus unterschiedlichen Themenbereichen wie Lexikographie, Linguistik, Verlagswesen, Forschung oder Softwareentwicklung. Andere Interessierte, die sich für die pädagogische, kulturelle, politische und soziale Bedeutung von Wörterbüchern im Alltag begeistern, sind ebenfalls willkommen.

Das Programm umfasst abwechslungsreiche Plenarvorträge, themenbezogene Sektionen, Softwarepräsentationen, Workshops vor Kongressbeginn und Vorstellungen von neuen Projekten und Nachwuchsarbeiten. Der EURALEX-Kongress ist ein ideales Forum für Diskussion und Austausch und bietet außerdem die Gelegenheit, gleichgesinnte Kolleginnen und Kollegen aus der ganzen Welt zu treffen. Der Kongress wird federführend vom Leibniz-Institut für Deutsche Sprache vom 12.-16. Juli 2022 in Mannheim organisiert.

EURALEX (European Association for Lexicography) ist der führende Berufsverband für alle, die in der Lexikographie und verwandten Bereichen arbeiten. In der sich schnell verändernden Welt der Sprachanalyse und -beschreibung bietet EURALEX ein internationales Forum für den Gedankenaustausch. EURALEX betreibt u. a. eine Mailingliste (euralex@freelists.org) zum Meinungsaustausch über verschiedenste lexikographische Themen. Die Liste steht allen interessierten Personen offen und ist nicht auf Mitglieder von EURALEX begrenzt.

Weitere Informationen zum Kongress finden Sie unter: https://euralex2022.ids-mannheim.de/de/startseite/

Deutscher
Terminologie-Tag e.V.

Deutsches
Institut für Terminologie e.V.

DTT-Fortbildung

© pixelio.de

DTT-Grundlagenseminar

„Terminologiearbeit – Grundlagen, Werkzeuge, Prozesse"

Referenten

- Prof. Dr. Petra Drewer
- Prof. Dr. Rachel Herwartz
- Prof. Dr. Klaus-Dirk Schmitz

Programm

- Terminologiearbeit – Grundlagen
- Werkzeuge Teil I: Recherche und Extraktion
- Methoden, Prozesse, Beteiligte, Qualifikation
- Werkzeuge Teil II: Management und Kontrolle

DTT-Vertiefungsseminar

„Terminologiearbeit – Projekte, Prozesse, Datenaustausch"

Referenten

- Beate Früh
- Dr. François Massion
- Dr. Detlef Reineke
- Dr. Annette Weilandt
- Angelika Zerfaß

Programm

- Prozesse und Projekte
- Einführung in Wissensorganisation
- Einführung in Datenaustausch
- Vertiefung Datenaustausch oder
  Vertiefung Wissensorganisation

Diese Veranstaltungen zählen zum **DTT-Terminologiezertifikat**.
Für den Erhalt des Zertifikats muss an einem DTT-Symposion, einem DTT-Grundlagenseminar, einem DTT-Vertiefungsseminar sowie an zwei DTT-Webinaren teilgenommen werden.

Termine für 2022 und weitere Informationen finden Sie in Kürze unter:
dttev.org/fortbildung

Stand: 1. Dezember 2021 | Änderungen vorbehalten.

DTT-Förderpreis 2021

Der diesjährige Förderpreis des DTT geht an Carola Maria Tremmel für ihre Bachelor-Arbeit zum Thema „Terminologiearbeit in einer geistlichen Gemeinschaft – Analyse des Ist-Zustandes und Entwicklung eines prozessorientierten Modells“.

Auf Grundlage der einschlägigen Fachliteratur erarbeitete Frau Tremmel für das Säkularinstitut der Schönstätter Marienschwestern einen praktikablen Terminologieprozess zur Erstellung mehrsprachiger Terminologie in einer Terminologiedatenbank. Mit ihrer Arbeit zeigte sie, dass Terminologieprozesse auch für die Anwendung außerhalb der klassischen Einsatzgebiete wie Industrie und Behörden modelliert werden können und eine optimistische Prognose erlauben.

Die Preisträgerin wird ihre Arbeit mit einem Vortrag auf dem nächsten DTT-Symposion im März 2023 genauer vorstellen.

Carola Maria Tremmel erhält den diesjährigen DTT-Förderpreis.

Die nächste Runde des DTT-Förderpreises läuft bereits. Folgende Termine sind wichtig:

Bis 30. Juni 2022:	Einreichung der Arbeit mit Begründung
Bis 30.September 2022:	Begutachtung und Bekanntgabe der Preisträger
Herbst/Winter 2022:	Abgabe eines Beitrags für den Tagungsband
Frühjahr 2023:	Vortrag auf dem DTT-Symposion

Weitere Informationen erhalten Sie unter: http://dttev.org/der-verband.html

DTT-Stammtisch

Nachdem die DTT-Lounge beim diesjährigen DTT-Symposion so gut ankam, haben wir uns dazu entschlossen, für unsere Mitglieder einen Stammtisch ins Leben zu rufen, der online stattfindet.

Der DTT-Stammtisch trifft sich seit September 2021 **an jedem ersten Mittwoch im Monat ab 20 Uhr** online.

Den Einladungslink erhalten alle DTT-Mitglieder per E-Mail an die bei uns in der Mitgliederdatenbank hinterlegte Adresse.

Für unsere Firmenmitglieder gibt es weiterhin das DTT-Industrieforum, an dem alle juristischen DTT-Mitglieder teilnehmen können, die weder Sprachdienstleister noch Beratungsunternehmen oder Toolhersteller sind. Bei Interesse am DTT-Industrieforum wenden sich sich bitte an **Tom Winter, winter@dttev.org**

Falls Sie DTT-Mitglied sind und schon länger nichts mehr vom DTT per E-Mail gehört haben, wenden Sie sich bitte an unsere Geschäftsführung und teilen Sie uns eine aktuelle E-Mail-Adresse mit.

DTT-Geschäftsführung
Dr. Annette Weilandt
geschaeftsfuehrung@dttev.org

DTT-Symposion 2023: Termin jetzt vormerken!

Das nächste DTT-Symposion findet vom
2. bis 4. März 2023 in Mannheim statt.

Terminology Summit 2022 in Iceland

Good news for everybody who loves terminology and languages: The FedTerm project consortium and TermNet, the International Network for Terminology, are going to organise a Pan-European Terminology Summit in Iceland in March 2022, within the FedTerm Project.

The Pan-European terminology summit will take place from 28-29 March 2022 in Iceland, as hybrid event.
The terminology summit will be a pan-European summit with showcases and best practice examples from Nordic countries. It is about terminology resources, management and terminology policies in so-called "lesser used languages". The focus will be on terminology in Nordic countries, such as Iceland, Greenland, Sweden, Denmark, the Baltic States, and the Gaelic language community in Ireland and the UK. However, Slovenian is also a "lesser used language", so that FedTerm partner JSI will present their national Slovenian project RSDO.

Further information:
https://termnet.eu/terminology-summit-2022-in-iceland

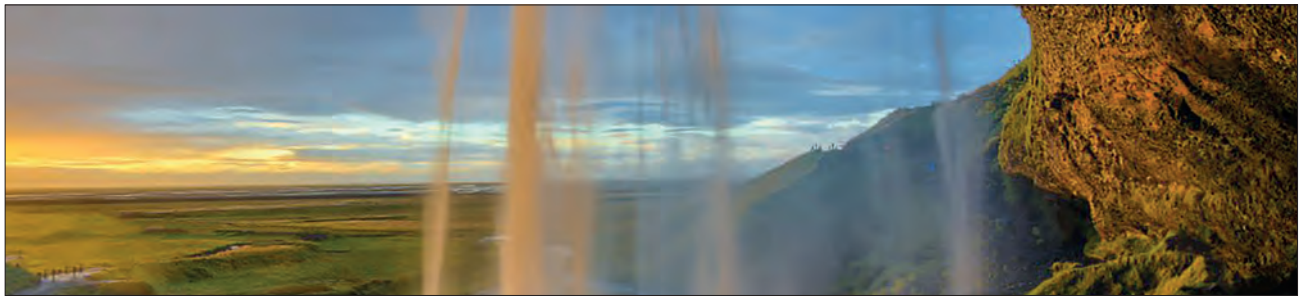


Fig. 1: Iceland (Source: termnet.eu | photographed by v2osk, unsplash.com)

1st International Conference On "Multilingual Digital Terminology Today"

Design, Representation Formats And Management Systems
16-17 June 2022, Padua, Italy

The professional responsible for designing and populating terminological resources is the terminographer. Producing new terminological resources requires eclectic research skills. The design and implementation phases involve a thorough preliminary analysis of the information needs of the potential user and an accurate assessment of the structural requirements of the resource.

In this context, the first international conference on "Digital terminology today. Design, representation formats and management systems" aims to bring together specialists in the disciplines of terminology, terminography, computational terminology, computational linguistics, NLP, in order to share methodological reflections on design approaches, representation formats and management systems of the digital terminology contained in the terminological resources.

In particular, the conference is organized around four main research topics:

- Topic 1: Analysis of the information needs of the future user of the resource
- Topic 2: Assessment of structural design approaches for terminological data collections
- Topic 3: Study of terminological metadata and data representation formats
- Topic 4: Methods of validating the ergonomics of a resource

The working languages of the conference are English and French.

Further information:
http://www.maldura.unipd.it/digital-terminology/en/

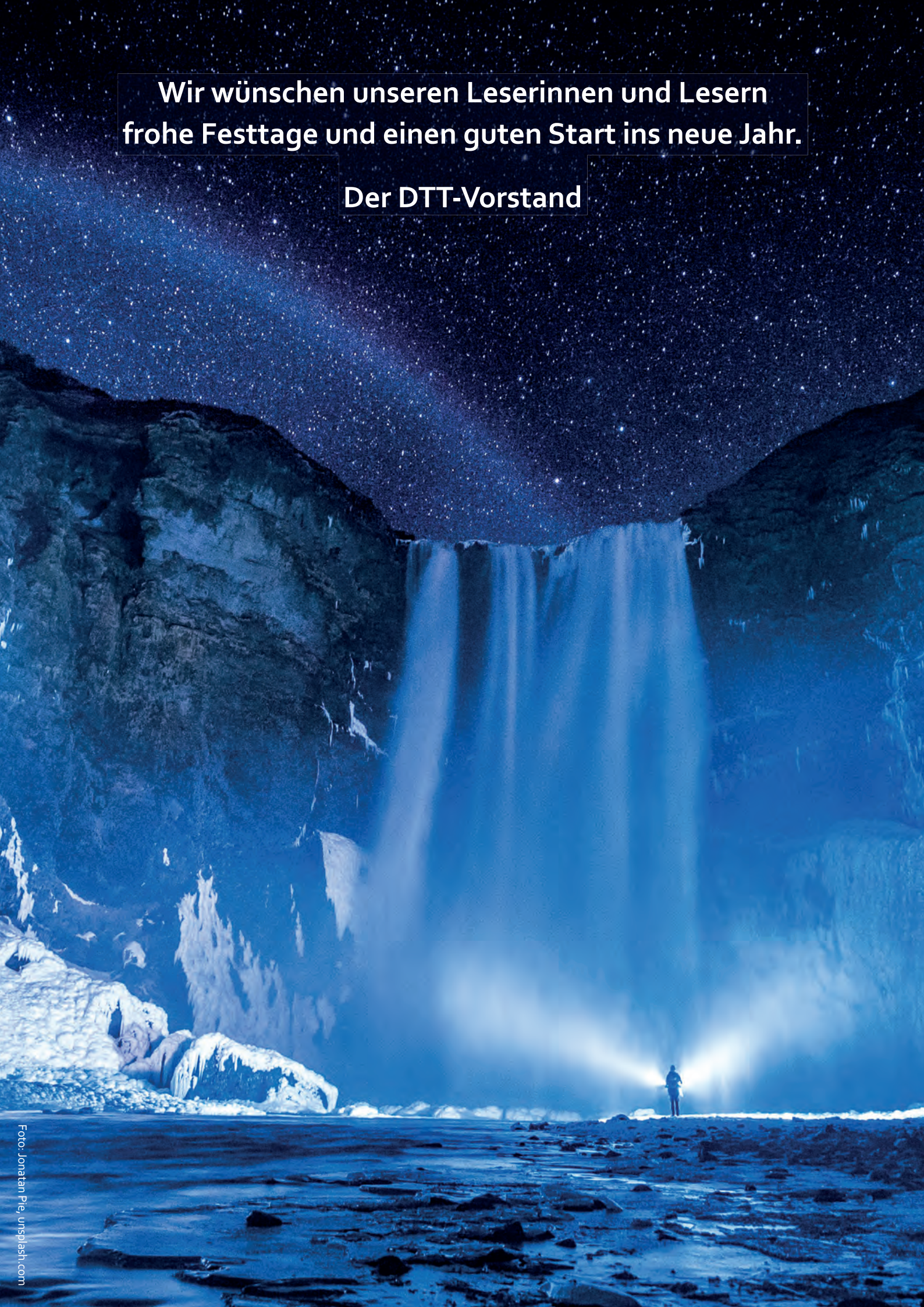
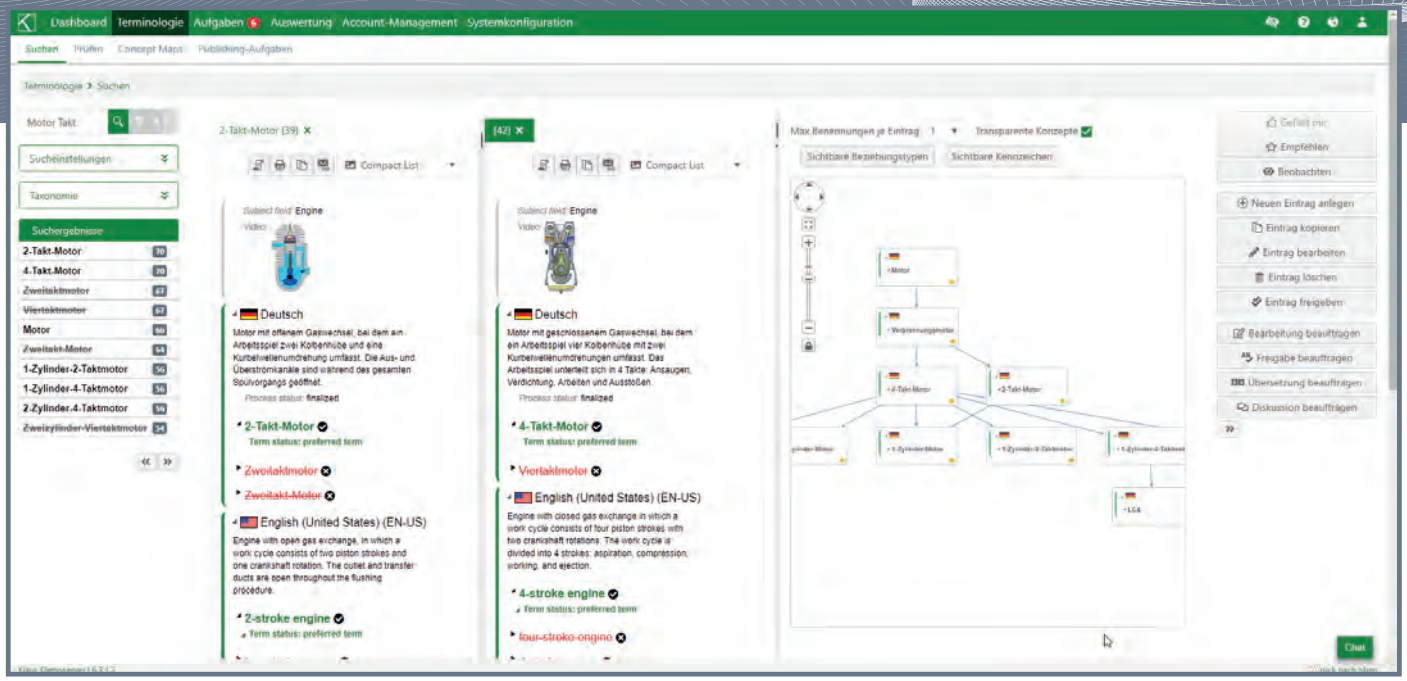


Foto: Jonatan Pie, unsplash.com

RWS KALEIDOSCOPE

ENTERPRISE TERMINOLOGY MANAGEMENT



- » **Unternehmensweiter Terminologiezugriff: Intuitiv, extrem anpassbar, informativ**
- » **Wissensmanagement mittels Concept Maps und Taxonomien**
- » **Kollaborative Workflows in allen Sprachen**
- » **Single Source of Truth - Die Termbank als Basis für andere Anwendungen wie CAT-Tools, MT, ERP, usw.**
- » **Terminologieprüfung bereits in der Ausgangssprache mit Plug-ins für MS Word, Adobe InDesign, XMetaL, Oxygen und Trados Studio**

QT KALCIUM Quickterm

CT KALCIUM Checkterm

Trados GmbH | Waldburgstr. 21 | 70563 Stuttgart
www.rws.com/de | trados-kontakt@rws.com

Kaleidoscope GmbH | Landstr. Hauptstr. 99 | 1030 Wien
www.kaleidoscope.at | info@kaleidoscope.at