Bernburg
Dessau
Köthen

Hochschule Anhalt
Anhalt University of Applied Sciences

Fachbereich 5
Informatik und Sprachen

Masterarbeit

Zur Erlangung des akademischen Grades
Master of Science (M. Sc.)

Yixin Xu

Vorname Nachname

Softwarelokalisierung, 4053969

Studiengang, Matrikelnummer

Thema:

Terminologieextraktion als Mittel zur Erstellung eines mehrsprachigen Terminologiebestands

Prof. Dr. Uta Seewald-Heeg

1. Prüfer/in

Dr. Horst Seiler

2. Prüfer/in

03.02.2015

Abgabe am

Selbstständigkeitserklärung

Hiermit erkläre ich, dass diese Arbeit von mir selbständig verfasst und in gleicher oder ähnlicher Fassung noch nicht in einem anderen Studiengang als Prüfungsleistung vorgelegt wurde. Ich habe keine anderen als die angegebenen Hilfsmittel und Quellen, einschließlich der angegebenen oder beschriebenen Software verwendet.

Ort, Datum	Unterschrift der Studierenden

Danksagung

Diese Masterarbeit wurde als Teil des Projektes „Terminologiemanagement – Schlüsselfaktor für Lokalisierungsprojekte und die Kommunikation international agierender Unternehmen“ am Fachbereich Informatik und Sprachen an der Hochschule Anhalt erstellt.

Ich möchte mich bei Frau Prof. Dr. Uta Seewald-Heeg für die Bereitstellung der Aufgabenstellung, die große Anzahl bereitgestellter hilfreicher Literatur, die Betreuung und die Korrektur dieser Arbeit an dieser Stelle herzlich bedanken.

Mein weiterer Dank gilt Herrn Dr. Horst Seiler für die Übernahme der Betreuung meiner Arbeit.

Bei Herrn Sebastian Hübel bedanke ich mich für die Übernahme der Korrektur dieser Arbeit.

Kurzfassung

Es gibt immer mehr Anforderungen an mehrsprachige Terminologie, um richtige und identische Übersetzungen von verschiedenen Informationsmaterialien zur Verfügung zu stellen. In diesem Rahmen wurde das Projekt „Terminologiemanagement Schlüsselfaktor für Lokalisierungsprojekte und die Kommunikation international agierender Unternehmen“ im Juli 2014 am Fachbereich Informatik und Sprachen an der Hochschule Anhalt eingerichtet. Terminologieextraktion ist der erste Schritt zur Verwaltung von Terminologie. Sie ist die Voraussetzung für eine richtige Sortierung verschiedener Synonyme eines Begriffs aus den vorhandenen Übersetzungen. Auf der Basis von Terminologieextraktion können die hochschulbezogenen Termini effizienter verwaltet werden. Diese Arbeit ist ein Teil von diesem Projekt, ein Pilotprojekt „Terminologieextraktion“. In diesem Projekt werden die Termkandidaten aus 47 Formularen der Hochschule Anhalt durch verschiedene Methoden extrahiert. Die extrahierten Termkandidaten werden dann in eine MultiTerm Datenbank importiert und bearbeitet. Durch Vergleich der Ergebnisse der extrahierten Termkandidaten werden die Vorteile und Nachteile der Extraktionsmethoden analysiert. Die Probleme und Lösungen bei der Bearbeitung von Termkandidaten werden gleichzeigt vorgestellt.

Abstract

There are more and more demands on the multilingual terms to provide correct and identical translations for informational materials. In this framework, a project “terminology management as key factor in localization projects and communicate international company” was established in the Department computer science and languages at the Anhalt University of Applied Sciences in July 2014. Terminology extraction is the first step to manage terminology. It is the basis for correct sorting of various synonyms of a term from the previous translations and efficient management of university-related terms. This work is a part of this project, a pilot project "terminology extraction". In this project, the term candidates will be extracted by different methods from 47 forms of HSA. The extracted term candidates are then imported and edited in a MultiTerm database. By compar-

ing the results of the extracted term candidates, the advantages and the disadvantages of the methods are analyzed. The problems and solutions in the processing of editing the term candidates are presented.

Abbildungsverzeichnis

Inhaltsverzeichnis

1 Einleitung

Termininologie wird nicht nur im Bereich der Übersetzung genutzt, sondern auch im gesamten Entwicklungs- und Lokalisierungsprozess eines Produktes. Ingenieure müssen bei der Entwicklung der neuen Produkte ihre Ergebnisse benennen und diese Namen in Unternehmensinformationsquellen eingeben. Logistikmitarbeiter verwenden Terminologie bei der Bestellung von Waren. Mitarbeiter in Dokumentationsabteilungen verwenden Terminologie, um technische Details von Produkten zu beschreiben. In der täglichen Korrespondenz ist die Verwendung von konsistenter Terminologie für Manager und Sekretäre sehr wichtig. Verkäufer verwenden Terminologie, um Wettbewerbsvorteile ihrer eigenen Produkte hervorzuheben. Terminologie kann zu einer reibungslosen Kommunikation zwischen Support- oder Kundendienstmitarbeiter und Kunden beitragen. Um Probleme mit Maschinen besser zu klären, sollen Mitarbeiter in der Fertigung die konsistente Terminologie verwenden. Bei der Bestellung von Teilen oder Produkten brauchen Verkäufer die Terminologie. Übersetzer verwenden Terminologie bei der Produktlokalisierung[1].

Terminologiemanagement umfasst alle Behandlungen der Terminologie. Durch die exakten Definitionen und die korrekte Bestimmung von Benennungen können nicht nur die Sprach- und Kulturbarrieren vermieden, sondern auch die konsistente Kommunikation inderhalb eines Unternehmen oder einer Organisation bzw. zwischen verschiedenen Unternehmen in einem Fachgebiet gewährleistet werden. Es gibt unterschiedliche Ziele in der Terminologieextraktion. Die Anwendungsfälle der verschiedenen Zielsetzungen werden im Buch „Einführung in die Terminologiearbeit[2]“ wie folgt beschrieben:

- Aufbau oder Ergänzung eines Terminologiebestands
- Aufwandsabschätzung und Vorbereitung eines Übersetzungsprojektes
- Überprüfung der terminologischen Konsistenz von Texten

[1] [Höge 2005]

[2] [Arntz 2014:Seite 244-245]

Der erste Fall trifft für diese Arbeit zu. Um einen mehrsprachigen Terminologiebestand zu erstellen, wird Terminologieextraktion als Mittel und auch als Grundlage dafür verwendet.

1.1 Motivation und Ziel der Arbeit

Mit den zunehmenden Anforderungen an englischsprachigen Studienangeboten, Informationsmaterialien, Zeugnissen usw. erhöht sich die Zahl der Übersetzungsaufträge an der Hochschule Anhalt (HSA). Dazu ist eine effiziente Verwaltung von mehrsprachiger Terminologie notwendig, um die Qualität der Übersetzung zu gewährleisten. Einige Termini in den vorhandenen Übersetzungen werden von verschiedenen Übersetzern unterschiedlich bezeichnet. Als Ergebnis nimmt die Zahl der Synonyme deutlich zu. Um solche Synonyme besser zu sortieren, die bevorzugten Benennungen leichter auszuwählen, und die mehrsprachige Terminologie einfacher zu verwalten, ist die Erstellung eines mehrsprachigen Terminologiebestands erforderlich.

Das Ziel dieser Arbeit ist, durch die Untersuchungen eine bessere Methode der Terminologieextraktion zu finden und dadurch eine Termbank für das Terminologiemanagement zu erstellen. Diese Termbank bietet zumindestet ein besseres Nachschlagen der Bedeutungen der Termini und der Übersetzungen vom Deutschen ins Englische. Damit Mitarbeiter der Abteilung Studentische Angelegenheiten (ASA) und andere hochschulinterne Mitarbeiter und Studierende auf den zu erstellenden Terminologiebestand zugreifen können, soll diese Termbank mit einem Termbank-Zugriff über einen Web-Browser online zur Verfügung stehen.

Terminologieextraktion ist die Grundlage zur Erstellung eines mehrsprachigen Terminologiebestands. In dieser Arbeit wird ein Terminologiebestand aus 47 PDF-Dateien, die Formulare der HSA sind, mit verschiedenen Methoden und Werkzeugen erstellt. Als Ergebnis werden alle Termini mit dem Terminologieverwaltungssystem SDL MultiTerm 2014 aufgebaut und zugänglich gemacht. Der Aufbau erfolgt mit SDL MultiTerm 2014 Desktop, die Bereitstellung für Mitarbeiter von ASA über die Browser-basierte Variante SDL MultiTerm 2014 Online.

Zur Erfassung der englischen und chinesischen Termini wird das DAAD-Wörterbuch und ein Deutsch-Chinesisches Universitätswörterbuch verwendet.

Im Einzelnen sind folgende Aufgaben durchzuführen:

- Alle wichtigen Termini werden extrahiert.
- Alle Benennungen eines Begriffs aus den verfügbaren Dateien sollen korrekt zusammengestellt werden.
- Die Bezeichnungen bzw. die Reihenfolge des Auftretens der beschreibenden Felder für den jeweiligen Terminus sollen identisch sein.
- Alle Synonyme in der Termbank sollen problemlos in SDL MultiTerm 2014 Desktop importiert oder sortiert werden.
- Alle Quellen eines Terminus, die aus den Formularen der HSA herangezogen werden, müssen in der Termbank angezeigt werden.
- Am Ende wird der Terminologiebestand in SDL MultiTerm Online (www.inf.hs-anhalt.de/multiterm/) zur Verfügung gestellt.

1.2 Aufbau dieser Arbeit

Das Kapitel zwei behandelt die Terminologieextraktion. Dieses Kapitel enthält die Parameter der einsprachigen und zweisprachigen Terminologieextraktion. Weiterhin stehen die Kriterien, die Methoden und die Techniken der Terminologieextraktion zur Verfügung, die Grundlage dieser Arbeit sind.

Im dritten Kapitel werden die Anweisungen zur Erstellung eines mehrsprachigen Terminologiebestands beschrieben. Von der Vorbereitung bis zur Bereitstellung eines Projektes wird Schritt für Schritt genau erklärt. Die Struktur der Datenbank wurde vor der Arbeit von Frau Prof. Dr. Uta Seewald-Heeg vorgegeben. So können die verschiedenen Datenbanken miteinander gut angepasst werden. Die Ergebnisse der Extraktion und der Inhalte der Termbank werden von der Projektkoordinatorin, Frau Prof. Dr. Uta Seewald-Heeg, strukturell überprüft, und darüber hinaus werden auch Verbesserungsvorschläge gegeben.

In Kapitel vier geht es darum, eine alternative Methode zur Terminologieextraktion mit Trados Studio vorzustellen. Dafür werden die Vorteile und Nachteile von Trados Studio mit einigen Beispielen genannt.

Im fünften Kapitel werden die Methoden der Terminologieextraktion evaluiert. Zusätzlich werden Hilfestellungen bei der Erfassung von Termkandidaten angeboten.

Im sechsten Kapitel wird diese Arbeit zusammengefasst. Die Besonderheiten der einsprachigen bzw. zweisprachigen Extraktion werden dargestellt. Außerdem gibt es auch weitere Empfehlungen zur Entwicklung und Optimierung der Extraktion und der Extraktionswerkzeuge.

2 Terminologieextraktion

In diesem Kapitel werden die Grundlagen der Terminologieextraktion vorgestellt. Dabei geht es im Wesentlichen um die Rolle von Terminologie in Unternehmen, um die Kriterien zur Term-Bereinigung und Term-Standardisierung und um die Methoden, Techniken bzw. Werkzeuge der Terminologieextraktion.

2.1 Die Rolle von Terminologie in Unternehmen

Terminologie ist die Gesamtheit der Begriffe und der Benennungen in einem Fachgebiet (ISO 2342). Massion präsentiert im Buch „Terminologiemanagement: Luxus oder Muss?[3]“ was extrahiert werden soll. **Benennungen und Kollokationen**, die sich eindeutig auf Produkte oder Leistungen des Auftraggebers beziehen, bzw. **allgemein bekannte Wörter**, die in Unternehmen eine besondere Bedeutung erhalten, werden aus den Textbeständen extrahiert. In der praktischen Arbeit werden auch viele Phrasen oder Sätze als firmenspezifische Termini in Unternehmen zusammengefasst, um die Ergebnisse der Übersetzung zu verbessern.

Um die terminologischen Daten in verschiedenen Sprachen besser zu verwalten, werden Terminologieverwaltungssysteme (TVS) entwickelt. Sie verfügen über zahlreiche für die TVS erforderliche Funktionen, wie zum Beispiel die Einstellungen von verschiedenartigen Eintragsstrukturen, die Schnittstellen zu Translation-Memory-Systemen oder Lokalisierungssystemen und die Möglichkeiten zum Datenaustausch in unterschiedlichen Formaten. Mit Hilfe von verschiedenen TVS können die Terminologiearbeiten effizient durchgeführt werden.

Auch beim Datenaustausch spielt Terminologie eine wichtige Rolle. Mit verschiedenen Methoden werden terminologische Daten importiert und exportiert. Viele Hersteller bieten das MultiTerm-Format für den Terminologieimport an. Das ist auch ein wichtiger Grund, warum SDL MultiTerm 2014 Desktop in dieser Arbeit verwendet wird.

[3] [Massion 2009]

Synonyme sollen beliebig bzw. unabhängig von dem Kontext austauschbar sein[4]. Synonyme werden meistens wegen der verschiedenen Übersetzungen von Fremdsprachen, der Ungenauigkeit von Definitionen oder der Verwendung von unterschiedlichen Stilen erzeugt. In der praktischen Arbeit werden die Beziehungen zwischen dem Oberbegriff und dem Unterbegriff häufig falsch oder ungenau differenziert. Dazu entstehen Quasisynonyme[5]. Sie werden heute in vielen Fällen als Synonyme verwendet. Die Behandlung von Synonymen und Quasisynonymen ist einer der Schlüsselpunkte bei der Terminologiearbeit. Die Einstellung vom beschreibenden Feld „Status“ mit zum Beispiel „bevorzugt, zugelassen oder verboten“ ist eine gute Lösung dafür.

2.2 Allgemeine Kriterien für die Term-Bereinigung

Das Ziel der Terminologieextraktion ist es, eine Terminologiedatenbank zu erstellen, deswegen müssen einige Prinzipien der Terminologiedatenbank auch hier bei der Extraktion beachtet werden, z. B. die Begriffsorientierung, die Benennungsautonomie und die Eindeutigkeit. Das heißt, die Termkandidaten müssen so extrahiert und bearbeitet werden, dass alle Kandidaten mit gleicher Bedeutung unter einem Modul oder einer Begriffsnummer sortiert werden sollen. Die Synonyme, Abkürzungen oder verschiedenen Schreibvarianten mit einer Begriffsnummer werden in gleiche Felder oder Zellen zusammen eingetragen. Und eine Benennung repräsentiert nur einen Begriff. Ihre anderen Bedeutungen werden mit anderen Begriffsnummern gekennzeichnet.

Die Anwendung der Terminologie spielt auch eine große Rolle bei der Term-Bereinigung. Beispielsweise für die Normierung dürfen nur firmenspezifische Fachwörter extrahiert werden. Aber für die Übersetzung können viele Kollokationen und allgemein bekannte Wörter bzw. Quasisynonyme und verschiedene Schreibvarianten extrahiert werden.

[4] [Seewald-Heeg 2011]

[5] Nach ISO 704 7.2.4 wird Synonymie in Synonymie (z. B. Studiengang mit NC und NC-Studiengang) und Quasisynonymie (z. B.Semester und Hochschulsemester) untergliedert. Synonymen sind beliebig austauschbar, während Quasisynonymen nur in bestimmten Kontexten austauschbar sind.

In „Modul 2 – Grundsätze und Methoden[6]“ der Publikation „Terminologiearbeit-Best Practices“ werden folgende Grundsätze für die Erfassung von Benennungen aufgelistet. Die Beispiele werden aus den Formularen der HSA genommen.

Die Benennungen sollen in folgender Form erfolgen:

- Ohne Artikel

Falsch	**Richtig**
einen NC-Studiengang	NC-Studiengang

- In der Grundform: Einzahl, Nominativ bzw. Infinitiv

Falsch	**Richtig**	**Ausnahme**
Studiengängen	Studiengang	Bewerbungsunterlagen
absolvierte	absolvieren	

- Groß- und Kleinschreibung wie im Fließtext

Falsch	**Richtig**	**Ausnahme**
Dualer Studiengang	dualer Studiengang	Kreative Kulturtechniken (Name eines Moduls)

- Natürliche Wortreihenfolge

Falsch	**Richtig**	**Ausnahme**
Prüfung, mündlich	mündliche Prüfung	Studium Generale (Name eines Moduls)

- Ohne Klammern oder andere Interpunktion in Benennungen

Falsch	**Richtig**	**Ausnahme**
Hochschulrektorenkonferenz (HRK)	Benennung 1: Hochschulrektorenkonferenz Benennung 2: HRK	Ethik & Ästhetik (Name eines Moduls)

[6] [Bauer 2014]

Es gibt Ausnahmen für jedes Kriterium. Für die Eigennamen oder beim Sonderfall wie zum Beispiel beim Auftreten in einer Softwareoberfläche oder als Name eines Moduls wird die Form der Benennung nicht geändert. Es ist unmöglich, dass hier alle Ausnahmen genannt werden. So ist es erfordlich, die Terminologen bei der Erfassung von Benennungen, mit Fachleuten oder Kunden zusammen zu diskutieren.

Müller hat die folgenden Kriterien zur Term-Bereinigung in ihrer Folie „Terminologielehre und Terminologieverwaltung[7]“ beschrieben (siehe Abbildung 2.1). Kriterien zur Term-Bereinigung sind unternehmensspezifisch. Solche Kriterien sind allgemein und bieten nur eine Richtung zur Term-Bereinigung. Die konkreten Probleme müssen genau analysiert werden. Die möglicherweise auftretenden Konflikte gegen die Kriterien müssen bei der Extraktion oder bei der Überprüfung mit Fachleuten zusammen behandelt werden.

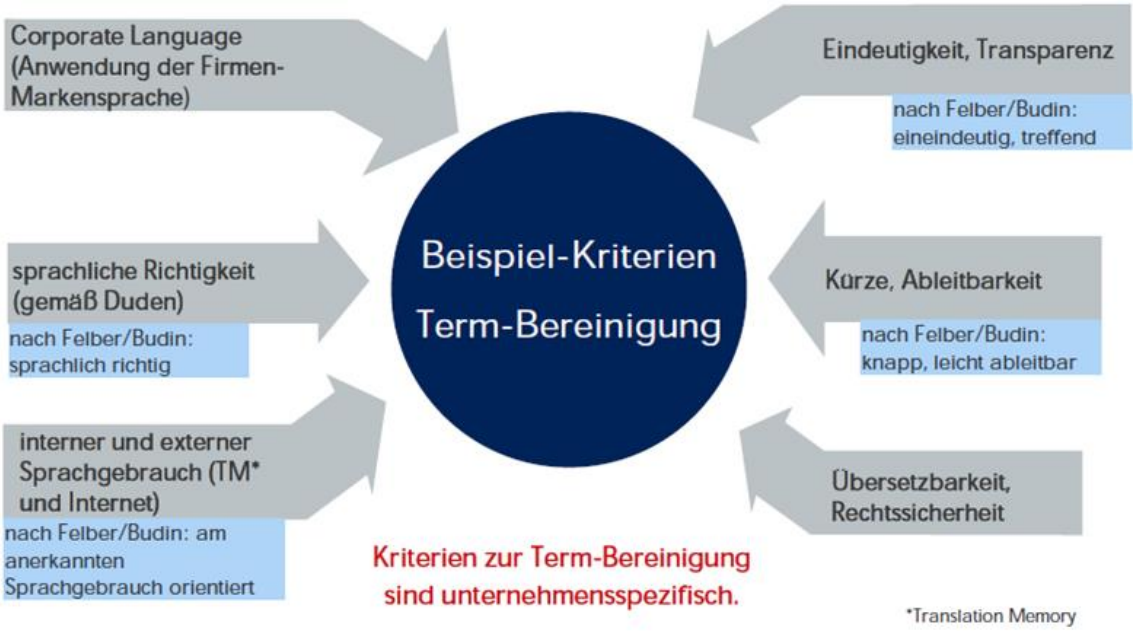


Abbildung 2.1 Kriterien zur Term-Bereinigung[8]

[7] [Müller 2014]

[8] Quelle: In Folie Terminologielehre und Terminologieverwaltung von Katja Müller

2.3 Terminologiearbeit durch Terminologieextraktion

Terminologieextraktion ist der erste Schritt der Terminologiearbeit. Terminologieextraktion ist der Teil der Terminologiearbeit, der darin besteht, Termini aus einem Korpus herauszufiltern (DIN 2342). Ein- oder mehrsprachige Terminologieextraktion hängt von den Ausgangsmaterialien ab. In der heutigen Forschung wird die mehrsprachige Terminologieextraktion häufig als zweisprachige Extraktion bezeichnet. Die Anforderungen an mehrsprachigen Terminologieextraktion erfolgt auf der Basis von ein- oder zweisprachiger Termextraktion. Später folgt bei der Terminologiearbeit noch die Erstellung eines Terminologiebestands und der Ausbau einer oder mehrerer Terminolgiebestände usw.

Durch verschiedene Verfahren können wichtige Wörter oder Fachtermini aus einem Text extrahiert werden. Das Gegenteil des Fachterminus ist das Stoppwort[9], das außer der Kalkulation von Worthäufigkeiten eine andere wichtige Unterstützung der Terminologieextraktion ist. Unter Stoppwort versteht man ein Wort, das bei der Termextraktion nicht beachtet wird. Eine Stoppwortliste enthält einige Stoppwörter, die in einem Text häufig aufgetretten und für die Erfassung eines Textes nicht relevant sind[10].

Vor der Extraktion der Terminologie müssen die folgenden Fragen gestellt werden: Was gehört zur Terminologie? Wozu braucht man Terminologieextraktion? Was gehört zu einem Fachterminus? Wie kann man Terminologie extrahieren? Welche Vorteile und Nachteile gibt es zwischen den verschiedenen Methoden der Terminologieextraktion? Worauf muss man bei der Terminologieextraktion achten? Wie sehen die Ergebnisse der Terminologieextraktion aus? Wie kann man die Methode bzw. die Technik weiter verbessern? Mit diesen Fragen startet diese Arbeit.

> *Terminologie kann auf verschiedene Arten gewonnen werden: durch manuelle oder maschinelle Extraktion aus vorhandenen Dokumenten des Unternehmens, durch Vorschläge bzw. Anfragen von Mitarbeitern und Kollegen aus den verschiedensten Abteilungen, durch Rückmeldun-*

[9] In MultiTerm Extract wird Stoppwort als Füllwort genannt.

[10] [IBM 2014]

gen in Bezug auf unklare Terminologie seitens der Übersetzer, durch automatisch protokollierte „erfolglose Suchen" in einer vorhandenen Terminologiedatenbank, durch Tools zur Autorenunterstützung, die automatisch neue Terminologiekandidaten sammeln, durch die systematische Erarbeitung eines Fachbereichs.[11]

Diese Arbeit verwendet die erste Art der Extraktion, die Extraktion aus vorhandenen Dokumenten, die auch am häufigsten bei Unternehmen genutzt wird.

2.3.1 Kriterien der Extraktion in diesem Projekt

Aufgrund der spezifischen Merkmale der Textsorte Antrag oder Fomular gibt es einige Besonderheiten. Die Termini in diesem Projekt werden nach den folgenden Kriterien als Grundlage extrahiert:

- Die Termini sollen so genau wie möglich (**Granularität**) extrahiert und bearbeitet werden. Damit können sie in zukünftigen Übersetzungen, z. B. in SDL Trados Studio mit der Funktion Terminologieerkennung, einfacher verwendet werden.

- **Die allgemeinsprachlichen Termini** werden nach Häufigkeit und Mehrdeutigkeit extrahiert, damit die weiteren Übersetzungen relativ leicht durchgeführt werden können.

- Die **Quasisynonyme** werden in dieser Arbeit extrahiert, damit sie später nach den spezifischen Anforderungen noch bereinigt oder bearbeitet werden können. Eine Anforderung ist, dass die entsprechenden Kontextbeispiele eingetragen werden müssen.

- Alle **Akronyme** werden extrahiert, obwohl manche nur dem allgemeinen Wortschatz zuzurechnen sind. (z. B. PF: Postfach)

- Die am häufigsten angewendete **Kombination** von Adjektiv und Substantiv wird im ersten Arbeitsschritt zusammen extrahiert. Dann werden

[11] [Arndt 2014: M5-13]

sie nach dem Kontext analysiert, ob sie getrennt werden oder zusammen beleiben können.

- **Mehrwortbenennungen**[12] werden bei der Extraktion mit Hilfe des Online-Wörterbuchs http://www.dict.cc/ nachgeschlagen und überprüft (z. B. abgeschlossene Berufsausbildung). Die Häufigkeit der Verwendung oder der Grad der Konsistenz der Übersetzungen (Art der Ausbildung) ist auch eine gute Referenz zur Extraktion von Mehrwortbenennungen.

66	Art d. Ausbildung	kind of training
67	Art d. Tätigkeit	kind of employment
67	Art der Tätigkeit	
68	Art der Arbeit	type of work

Abbildung 2.2 Extraktion von Mehrwortbenennungen

- Wortgruppen (**Mehrwortbenennungen mit Ellipse**) werden in zwei getrennte Benennungen untergliedert, z. B. wird die Bezeichnung „berufliche Ausbildung bzw. Tätigkeit“ bei der Extraktion in zwei Bezeichnungen „berufliche Ausbildung“ und „berufliche Tätigkeit“ aufgespalten. Eine Ausnahme ist beispielsweise die Bezeichnung eines Fachbereichs oder eine Modulbezeichnung wie „Elektro- und Informationstechnik“, die genauso wie im Text sein soll.

- Alle **Schreibvarianten**, z. B. die Verwendung von Bindestrichen, Ziffern, Zahlwörtern, Fugenelementen und Flexionen von Benennungen (Genitiv- und Dativbildung), der Umgang mit Abkürzungen, Groß- und Kleinschreibung, sollen extrahiert werden. Sie können später durch Eingabe eines Status wie „bevorzugt, zugelassen oder verboten“ begrenzt.

2.3.2 Kriterien zur Bewertung von Benennungen

Die Kriterien zur Term-Standardisierung oder zur Bewertung von Benennungen sind die Grundlage für die Einstufung vom Verwendungsstatus, d. h. diese Kri-

[12] Mehrwortbenennungen sind zusammenhängende Wortgruppen, deren Beständteile Leerzeichen getrennt sind, während Einwortbenennungen Simplizia und Komposita umfassen.

terien spielen eine entscheidende Rolle bei der Bestimmung, welche Benennung bevorzugt benutzt werden soll. Diese Kriterien gelten auch für die Bildung einer neuen Benennung. Im „Modul 3 - Benennungen[13]“ der Publikation „Terminologiearbeit – Best Practices“ werden verschiedene denkbare Kriterien zur Bewertung von Benennungen in Abbildung 2.3 angezeigt. Eine allgemeingültige Gewichtung der Kriterien ist nicht möglich. Es gibt noch Konflikte zwischen den Kriterien, z. B die Genauigkeit und die Eindeutigkeit gegen die sprachliche Ökonomie. Die Genauigkeit und die Eindeutigkeit einer Benennung und die sprachliche Ökonomie können nur schwer koexistieren. Je länger eine Benennung ist, desto genauer und eindeutiger ist ihre Bedeutung. Die Kurzform einer Benennung kann nicht alle Merkmale des Begriffs enthalten. Aber man neigt zur Verwendung der Kurzform, denen meistens die Genauigkeit und die Eindeutigkeit fehlt. Die Sonderfälle müssen spezifisch analysiert werden.

[13] [Drewer 2014]

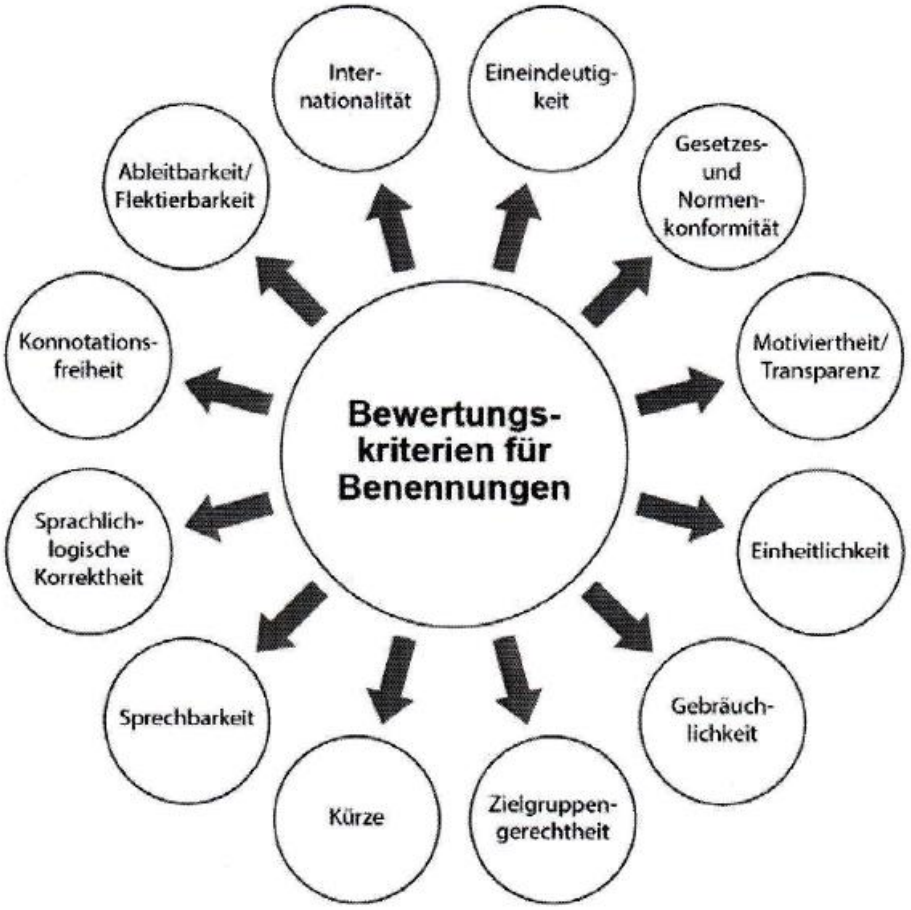


Abbildung 2.3 Bewertungskriterien für Benennungen[14]

Auf der Basis der Abbildung 2.3 werden die Krieterien wie folgt zusammengefasst:

Gesetzes- und Normenkonformität

Ist eine Benennung genormt oder in fachliche Verbände oder Gesetze vorgeschrieben, soll sie als Vorzugsbenennung bezeichnet werden.

Eineindeutigkeit

Eine eindeutige Benennung ist nur einem Begriff zuzuordnen, und als Vorzugsbenennung auszuwählen. Eine eindeutige Benennung ist meistens ein Unterbegriff in einem Begriffssystem. Z. B. im Satz „Bitte geben Sie Ihre Schulausbil-

[14] Quelle: Modul 3 – Benennungen der Publikation Terminologiearbeit – Best Practices

dung vom ersten Tag bis zum Abschluss, der Sie zur Aufnahme eines Hochschulstudiums im Land der Ausstellung berechtigt, an.“ bedeutet das Wort „Abschluss“ „Schulabschluss“ aber nicht „Studienabschluss“ oder „Hochschulabschluss“. Nach diesem Kriterium wird das Wort „Schulabschluss“ bevorzugt ausgewählt.

Transparenz/ Motivation

Eine transparente/ motivierte Benennung enthalt die entscheidenden Merkmale des Begriffs. Die Bedeutung einer Benennung wird durch ihre Teile deutlich erklärt, z. B. „Auswahlkriterium“ (Kriterium, nach dem jemand, etwas ausgewählt wird.) Je länger ein Terminus ist, desto transparenter ist das Wort.

Einheitlichkeit

Wenn eine Benennung zu einem bestimmten Begriffssystem gehört oder eine ähnliche Form mit einem vorhandenen Terminus hat, ist sie bevorzugt im Vergleich zu ihren Synonymen. Folgendes Beispiel veranschaulicht das genauer:

Abgabetermin (Abgabedatum) – **Aufnahmetermin** – **Prüfungstermin** (Prüfungstag, Prüfungsdatum)

In den extrahierten Termkandidaten werden „Abgabedatum“ als Synonym von „Abgabetermin“ und „Prüfungstag und Prüfungsdatum“ als Synonyme von „Prüfungstermin“ eingetragen. Die drei Benennungen „Abgabetermin, Aufnahmetermin und Prüfungstermin“ haben das gleiche Grundwort „Termin“, deswegen sollen die drei bevorzugt ausgewählt werden.

Kürze

Kürze bezieht sich hier darauf, dass die Einwortbenennungen statt der Mehrwortbenennungen als bevorzugt ausgewählt werden. Z. B. in der Datenbank wird „**Prüfungsanmeldung**“ aber nicht „**Anmeldung zur Prüfung**“ als „bevorzugt“ eingetragen. Eine verkürzte Benennung verliert nach dem Auslassen des Wortbestandteils ihre Eindeutigkeit und Richtigkeit.

Gebräuchlichkeit

Die Gebrächtlichkeit bedeutet, die zu sehr fachspezifischen Benennungen zu vermeiden. Dieses Kriterium kann dem Kriterium „Gesetzes- und Nomenkonformität“ durchaus widersprechen, weil die genormten oder in fachlichen Verbänden oder Gesetzen vorgeschriebenen Benennungen fachspezifisch sind und einige davon selten gebraucht werden. In diesem Fall sind die Kriterien bei jedem Unternehmen oder jeder Organisation nach Anforderungen zu nutzen.

2.4 Methoden und Techniken der Terminologieextraktion

Bezüglich der Ausgangsmaterialien lässt sich die Terminologieextraktion in einsprachige und zweisprachige Extraktion klassifizieren. Die heutigen Terminologieextraktionen sind maschinengestützt. Neben der Identifizierung von Termkandidaten werden gleichzeitig einige relevante Zusatzinformationen wie die Kontextbeispiele und ihre Quellen extrahiert. Die vielfältigen Werkzeuge und Methoden bieten viele Möglichkeiten zur Terminologieextraktion an. Je einfacher die Verfahren verwendet werden, desto höher ist die Qualität. Einfache Verfahren sind aber auch zeitaufwändiger und können durch Übermüdung oder wegen fehlender Kenntnisse in bestimmten Fachgebieten Fehler erzeugen. Durch Extraktionswerkzeuge wird Zeit eingespart. Aber unter Berücksichtigung der Vollständigkeit und der Richtigkeit entstehen gleichzeitig Nachteile. Für alle Extraktionen ist die menschliche Nachbearbeitung erforderlich.

2.4.1 Einfache Verfahren

Zweifellos ist die manuelle Terminologieextraktion die einfachste Methode. Die Termkandidaten werden von einem Terminologe in einem Artikel manuell markiert und herausgefiltert. Und dann werden sie schriftlich festgehalten oder in einem computergestützten Werkzeug z. B. in eine Excel-Tabelle eingetippt. Gleichzeitig sollen auch die Zusatzinformationen wie die Kontextbeispiele und die Quellenangaben hinzugefügt werden.

Eine andere einfache Methode zur Extraktion ist die Anwendung von einem Textverarbeitungsprogramm wie zum Beispiel MS-Word und MS-Excel. Alle

Leerzeichen und Interpunktionszeichen eines Textes werden durch die Absatzmarke (^p) in MS-Word ersetzt. In Excel werden die Duplikate[15] und Stoppwörter gelöscht, unterschiedliche Wortformen in ihre Grundform gewandelt, und die Synonyme zusammen eingeordnet.

Mit der Funktion „Ersetzen“ werden einzelne Bestandteile der Mehrwortbenennungen in Excel voneinander getrennt, so dass die Mehrwortbennungen nicht mehr erkannt werden können. Um dieses Problem zu lösen, kann die in MS-Word erzeugte Wortliste, die in Abbildung 2.4 angezeigt wird, stufenförmig in Excel hinzugefügt werden. Wieviele Spalten erstellt werden sollen, hängt von der maximalen Anzahl der einzelnen Wörter der Mehrwortbenennungen ab. Beispielsweise besteht eine der längsten Mehrwortbenennungen im Text „Anhalt University of Applied Sciences“, aus fünf Wörtern. In Excel werden entsprechend fünf Spalten erzeugt. Die Kopie der Wortliste wird ab der fünften Zeile in die erste Spalte eingefügt. In Abbildung 2.4 werden die Mehrwortbenennungen bzw. Eigennamen „Hochschule Anhalt“, „Anhalt University of Applied Sciences“ als Beispiele dargestellt. Es ist schwer, alle Mehrwortbenennungen zu finden, deswegen muss die Ausgangsdatei immer überprüft werden. Ein weiteres Beispiel ist hier die Bezeichnung des Fachbereichs „Architektur, Facility Management und Geoinformation“. Da das Komma nach dem Wort „Architektur“ in MS-Word gelöscht wird, ist dieser Mehrwortbenennung ohne Vorwissen oder Vorschau des Ausgangstextes sehr schwer zu erkennen.

Die einfachen Verfahren werden häufig als manuelle Terminologieextraktion bezeichnet.

[15] In Excel gibt die Funktion „Duplikate entfernen“ unter Menüleiste „Daten“

1					Hochschule
2				Hochschule	Anhalt
3			Hochschule	Anhalt	Anhalt
4		Hochschule	Anhalt	Anhalt	University
5	Hochschule	Anhalt	Anhalt	University	of
6	Anhalt	Anhalt	University	of	Applied
7	Anhalt	University	of	Applied	Sciences
8	University	of	Applied	Sciences	Fachbereich
9	of	Applied	Sciences	Fachbereich	Architektur
10	Applied	Sciences	Fachbereich	Architektur	Facility
11	Sciences	Fachbereich	Architektur	Facility	Management
12	Fachbereich	Architektur	Facility	Management	und
13	Architektur	Facility	Management	und	Geoinformation
14	Facility	Management	und	Geoinformation	
15	Management	und	Geoinformation		
16	und	Geoinformation			Martin-Luther-Ur
17	Geoinformation			Martin-Luther-	Halle-Wittenberg
18			Martin-Luther-	Halle-Wittenbe	Philosophische
19		Martin-Luther-	Halle-Wittenbe	Philosophische	Fakultät
20	Martin-Luther-	Halle-Wittenbe	Philosophische	Fakultät	I
21	Halle-Wittenbe	Philosophische	Fakultät	I	
22	Philosophische	Fakultät	I		
23	Fakultät	I			

Abbildung 2.4 Mehrwortbenennungen extrahieren

2.4.2 Nutzung von Konkordanzprogrammen

Die Konkordanzprogramme bieten die einsprachige Extraktion an, und bestehen meistens nur aus dem TXT oder RTF Format[16]. Das Konkordanzprogramm wird in „Modul 4 – Werkzeuge und Technologien[17]“ in „Terminologiearbeit – Best Practices“ wie folgt beschrieben:

Ein Konkordanzprogramm:

- erstellt eine Liste **aller** Benennungen (bestehend aus 1 bis n Wörtern),
- zeigt die Häufigkeit jeder Benennung an,
- zeigt den Kontextsatz einer Benennung an,
- ermöglicht das Anlegen von Stoppwortlisten[18],

[16] [Zerfass 2008]

[17] [Ferrari 2014]

[18] Die gestrichenen Wörter werden in die Stoppwortliste aufgenommen und beim nächsten Extraktion nicht noch einmal angezeigt.

- ist sprachunabhängig,
- und exportiert die Ergebnisse (Liste der Benennungen/ Kontextsätze) meist in ein tabulatorgetrenntes Format.

In Abbildung 2.5 werden die Beispiele von Konkordanzprogrammen aus der Präsentation „Terminologie Management – Methoden und Programme zur Erfassung, Bearbeitung/ Verwendung und Prüfung von Terminologie[19]“ von Angelika Zerfass in 2008 dargestellt. Die Extraktionswerkzeuge Simple Concordance Program (SCP) (http://www.textworld.com) und ExtPhr32, (http://publish.uwo.ca/~craven/freeware.htm) sind Beispiele des Einsatzes von Konkordanzprogrammen. Einige der Konkordanzprogramme bieten zusätzlich eine Lemmatisierung oder Normalisierung an, damit die Termkandidaten auf ihre Grundform zurückgesetzt werden. Obwohl Mehrwortbenennungen manchmal nicht erkannt werden, können sie von Terminologen mithilfe von einem KWIC-Index (Keyword in Context) identifiziert werden[20]. In einem KWIC-Index werden die Bestandteile einer Mehrwortbenennung hervorgehoben, dadurch wird diese Mehrworbenennung bei der Bearbeitung schnell erkannt.

[19] [Zerfass 2008]

[20] [Arntz 2014]

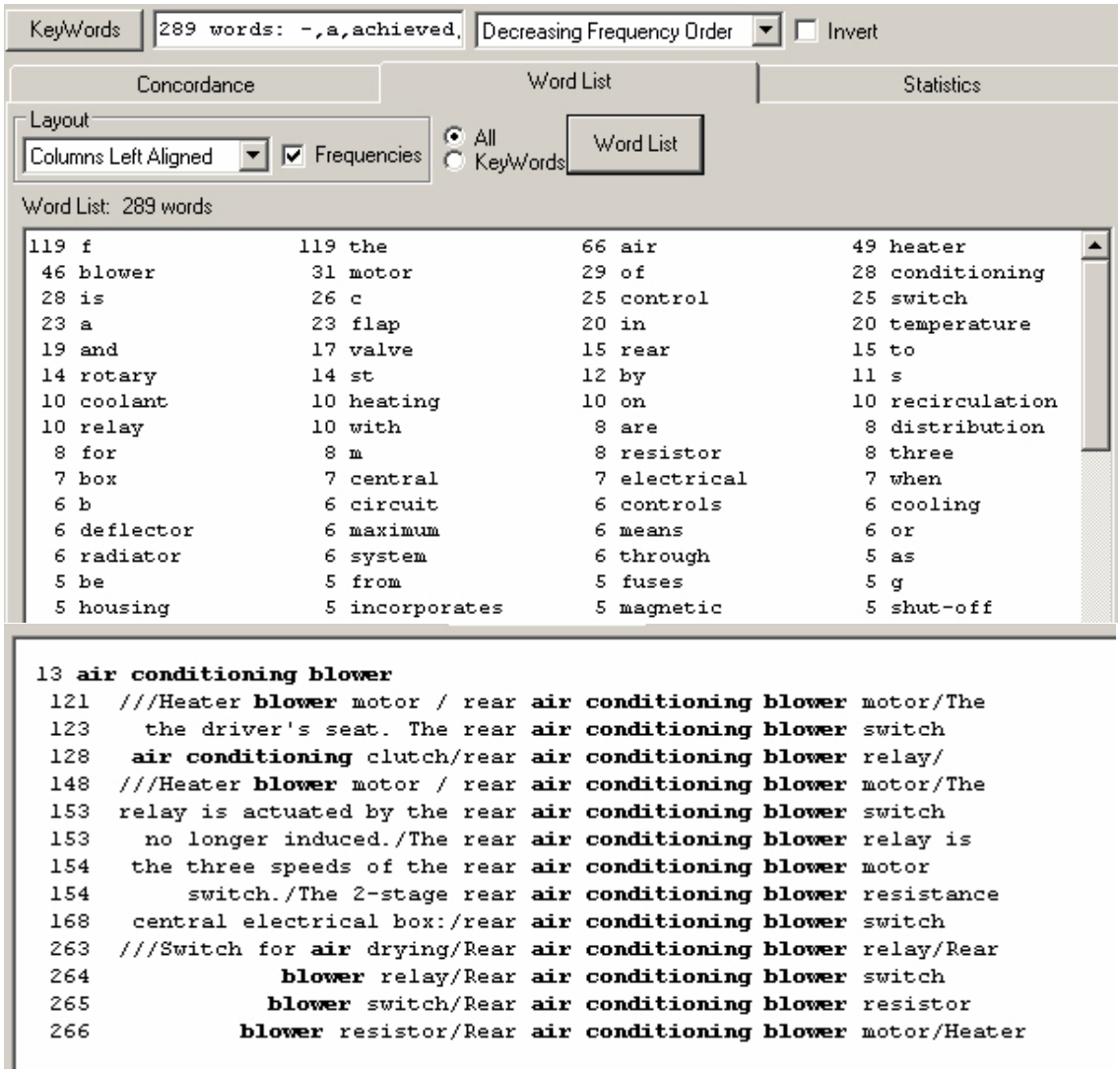


Abbildung 2.5 Konkordanzprogramme aus Zerfass 2008

2.4.3 Statistische Verfahren

Ein statistisches Terminologieextraktionsprogramm funktioniert durch die Analyse der Häufigkeit eines Wortes. Häufig auftretene Wörter werden als Termkandidaten gekennzeichnet. Grundsätzlich wird der Wert der Auftretenshäufigkeit durch Einstellen von „Silence“ oder „Noise“ geändert. Bei Silence ist die Qualität der Termkandidaten besser, aber die Zahl der Termkandidaten ist sehr gering. Im Gegensatz dazu gibt es bei Noise viele falsche oder irrelevante Termkandidaten, welche manuell gelöscht werden müssen. Es ist sinnvoll, bei verschiedenen Texten bzw. Textsorten durch Untersuchungen einen optimalen Wert zu finden. Durch das Hinzufügen einer Stoppwortliste nimmt die Qualität der Extraktion zu. Mit statistischen Verfahren entsteht ein- oder zweisprachige Extraktion. Ein deutliches Merkmal ist, dass die Arbeit unabhängig von den

Sprachen ist. Bei einem statistischen Terminologieextraktionsprogramm gibt es die Möglichkeit, die Kontextbeispiele anzuzeigen. Die heutzutage häufig verwendeten Terminologieextraktionsprogramme werden meistens mit statistischen Verfahren erzeugt, z. B. MultiTerm Extract, Déjà Vu Lexicon, Heartsome Dictionary Editor, across, TermiDOG (www.dog-gmbh.de) und Chamblon Terminology Extractor (http://www.chamblon.com/terminologyextractor.htm). Die häufiger in gleicher Reihenfolge zusammen auftretenden Wörter werden als Mehrwortbenennungen gekennzeichnet. Obwohl dieses Verfahren viele Mehrwortbenennungen identifizieren kann, fehlt es leider ein bisschen an Korrektheit.

2.4.4 Linguistische Verfahren

Wie der Name sagt, arbeiten linguistische Verfahren stark sprachabhängig. Nach den in der Terminologie üblichen Wortbildungsmustern einer Sprache werden die Besonderheiten von Benennungen dieser Sprache definiert. Bei den linguistischen Verfahren werden verschiedene Benennungsvarianten identifiziert. Die verschiedenen flektierten Wortformen werden automatisch auf ihre Grundformen zurückgeführt, dadurch wird der Nachbearbeitungsaufwand verringert. Wegen der unterschiedlichen Eigenschaften von Sprachen unterstützen die Terminologieextraktionsprogramme mit linguistischen Verfahren nur eine oder wenige Sprachen. Die Extraktionswerkzeuge mit linguistischen Verfahren sind beispielsweise Synthema Terminology Wizard (http://www.synthema.it/index.php/en/Prodotti/terminologywizard/Terminology-Wizard.html) und der SDL PhraseFinder.

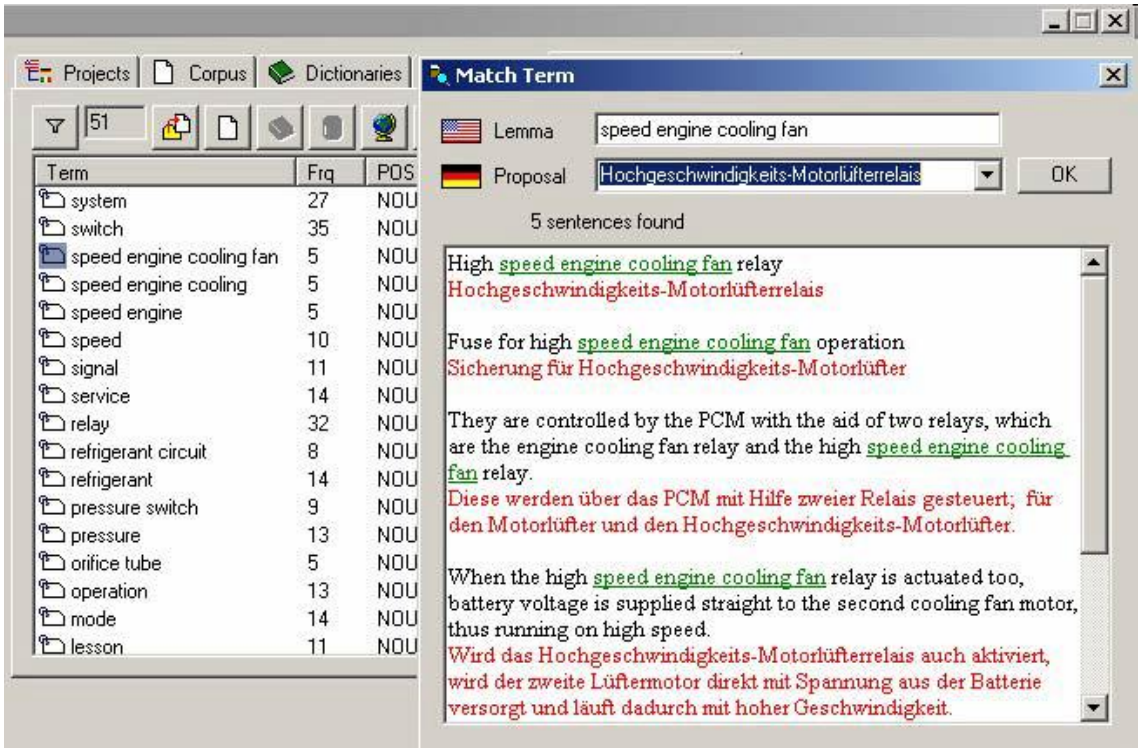


Abbildung 2.6 Linguistische Extraktion aus Zerfass 2008

2.5 Werkzeuggestützte Terminologieextraktion

Massion hat viele Terminologieprogramme mit Extraktionsfunktion im Buch „Terminologiemanagement – von der Theorie zur Praxis[21]“ bzw. die Folie „Terminologiemanagement_FH_Anhalt-Massion-2014_Teil_2[22]“ aufgelistet. In dieser Arbeit werden zwei Extraktionswerkzeuge, SDL MultiTerm 2014 Extract und memoQ, die an der Hochschule verfügbar sind, vorgestellt.

[21] [Massion 2009]

[22] [Massion 2014]

Produkt	Beschreibung	Adresse
acrolinx® Terminology Lifecycle Management	Terminologieextraktion und -verwaltung, Terminologieplattform	www.acrolinx.com
Across crossTerm	System für die Extraktion, Verwaltung und Prüfung von Terminologie.	www.across.net
AntConc3.2	Statistische Textanalyse, kostenlos	www.antlab.sci.waseda.ac.jp
CATS	Terminologieerfassung und Verwaltung. Prof. Schmitt, Uni Leipzig	www.cats-term.com
Concord	Wordsmith: Konkordanz und Keywords	www.lexically.net
Concordance	Einsprachige Terminologieextraktion mit Satzbeispielen. Preiswert.	www.concordancesoftware.co.uk
Extphr33	Einsprachige Terminologieextraktion, kostenlos	http://publish.uwo.ca/~craven/freeware.htm
Quickterm	Workflow-Tool für Terminologie	www.kaleidoscope.at
Heartsome Dictionary Editor	XML-basiertes Wörterbuch mit Extraktionsfunktion. Heartsome	www.heartsome.net
KWIC Concordance	Korpusstatistik, Konkordanz, kostenlos	www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html
KWICKWIC	Konkordanz	www.kwickwic.com
Linguo	Selbständiges Terminologieverwaltungsprogramm - Export nach TMX + Text - Integrierbar in Word.	www.lexicool.com
LookUp	Internetfähige Terminologieplattform - Kaufversion + Mietsoftware	www.dog-gmbh.de
MultiTerm und MultiTerm Extract 9	Von SDL/Trados angeboten. Integriert in Trados Technologie. Mehrsprachig.	www.sdl.com
qTerm	Webfähiges System für die Verwaltung von Terminologie.	www.kilgray.com
SDLX Phrase Finder	Terminologieextraktion mit linguistischen Fähigkeiten. Für sieben Sprachen verfügbar.	www.sdl.com
Similis	Automatisches bilinguales Extraktionsprogramm. Gute Ergebnisse	www.lingua-et-machina.com
Simple Concordance Program (SCP)	Konkordanz- und Extraktionsprogramm (einsprachig) - Freeware	http://www.textworld.com/scp/
Synchroterm	Zweisprachige Terminologieextraktion semi-automatisch, effizient	www.terminotix.com / www.dog-gmbh.de
Terminology Extractor	Einsprachige Extraktion von Terminologie	www.chamblon.com/terminologyextractor.htm
TermStar	Terminologiemodul von Star Transit	www.star-ag.ch
Termweb	Terminologieverwaltungssystem	www.interverbum.com
termXplorer	Terminologieverwaltungssystem	www.termxplorer.com
Textanz	Einsprachige Terminologieextraktion mit Kontextangabe	www.cro-code.com
Textstat	Statistische Textanalyse, kostenlos	www.niederlandistik.fu-berlin.de/textstat/
Tippyterm	Add-on zu MS-Office-Anwendungen, editierbares Wörterbuch	www.syskon.com
Webterm	Internetfähige Terminologieplattform von Star AG	www.star-ag.ch

Abbildung 2.7 Extraktionswerkzeuge[23]

2.5.1 SDL MultiTerm 2014 Extract

SDL MultiTerm 2014 Extract bietet die einsprachige und zweisprachige Extraktion an. Außerdem gibt es noch viele zusätzliche Funktionen zur Terminologieextraktion, z. B. die Möglichkeit des Hinzufügens einer externen Stoppwortliste[24] . Die Einstellungen während der Extraktion sind sehr flexibel. Bei SDL MultiTerm 2014 Extract gibt es noch die Funktion manuelle Extraktion aus einem Dokument (siehe Abbildung 2.8). Der KWIC-Index liegt entweder im Feld „Sätze generieren“ oder im Fenster „Konkordanz“. Je größer die Dateien, desto besser wird die maschinelle Terminologieextraktion funktionieren, weil es ein besseres statistisches Ergebnis dafür anbietet.

[23] Quelle: Folie von Dr. Massion „Terminologiemanagement_FH_Anhalt-Massion-2014_Teil_2“

[24] Füllwörterliste in SDL MultiTerm 2014 Extract

Am Anfang sind fünf Projekttypen auszuwählen. In dem Projekt „einsprachige Termextraktion“ und „zweisprachige Termextraktion“ können die Termkandidaten aus einsprachigen und zweisprachigen Dokumenten ausgelesen werden. In „Übersetzungsprojekt“ können vorhandene Termbanken mit neuen Übersetzungen für bereits in MultiTerm gespeicherte Termini aktualisiert[25] werden. In „Wörterbucherstellungsprojekt“ können zweisprachige Wörterbücher erstellt werden. Nach der Anzahl der Sprachen in dem Ausgangsmaterial wird die „einsprachige Termextraktion“ oder „zweisprachige Termextraktion“ ausgewählt.

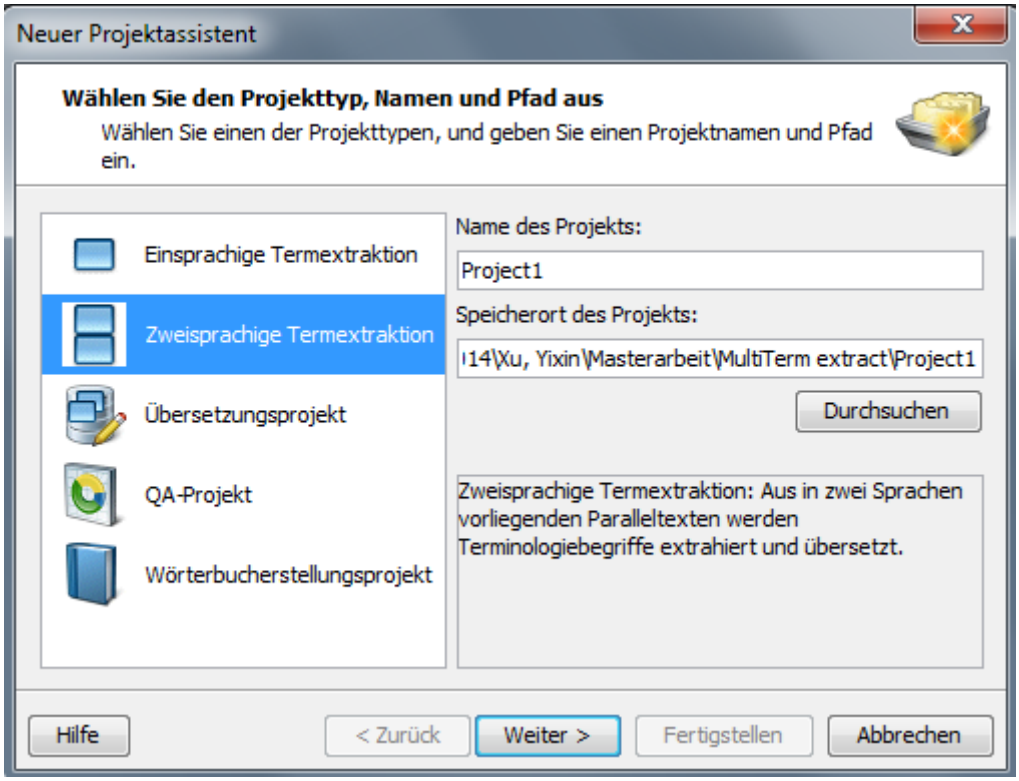


Abbildung 2.8 SDL MultiTerm 2014 Extract - Projekttyp auswählen

Es ist möglich SDL MultiTerm 2014 Extract mit SDL MultiTerm 2014 Desktop zu verbinden. Bei der Auswahl der Termbank und der Sprachen (siehe Abbildung 2.9) muss zuerst eine Termbank ausgewählt werden, worin die bestätigten Termkandidaten nach der Extraktion exportiert werden sollen. „Keine Termbank“ bedeutet, dass eine mit Tabulator getrennte Termliste nach der Extraktion in TXT-Format exportiert wird.

[25] [SDL MultiTerm 2014 Extract]

Abbildung 2.9 SDL MultiTerm 2014 Extract - Termbank und Sprachen auswählen

SDL MultiTerm 2014 Extract unterstützt viele Dateiformate. Beim Hinzufügen von Dateien müssen die entsprechenden Dateiformate ausgewählt werden. Abbildung 2.10 zeigt alle unterstützten Dateiformate bei der einsprachigen und zweisprachigen Termextraktion an.

Unterstützte Dateiformate

Einsprachiges Projekt zur Termextraktion:

- TXT – Nur-Text
- RTF – Rich Text Format
- DOC – Word-Dokumente
- HTML, HTM, JSP, ASP, ASPX
- SGML, SGM, XML
- Ventura (*.txt)
- PageMaker (*.txt)
- QuarkXPress (*. qsc, *.txg, *.ttg, *.tag)
- TTX – TRADOStag-Dateiformat
- InDesign (*.isc)
- PowerPoint-Dateien (*.ppt, *.pps, *.pot)
- Excel-Dateien (*.xls, *.xlt)

Zweisprachiges Projekt zur Termextraktion:

- TXT – Nur-Text
- RTF – Rich Text Format
- DOC – Word-Dokumente
- HTML, HTM, JSP, ASP, ASPX
- SGML, SGM, XML
- Ventura (*.txt)
- PageMaker (*.txt)
- QuarkXPress-Dokumente (*.qsc; *.txg; *.ttg; *.tag)
- TTX – TRADOStag-Dateiformat
- TMX – Translation Memory-Austauschformat
- TMW – Translator's Workbench
- Translator's Workbench-TXT-Format
- Translator's Workbench-RTF-Format
- TTX – TRADOStag-Dateien
- InDesign (*.isc)
- PowerPoint-Dateien (*.ppt, *.pps, *.pot)
- Excel-Dateien (*.xls, *.xlt)

Abbildung 2.10 SDL MultiTerm 2014 Extract - Unterstützte Dateiformate[26]

In SDL MultiTerm 2014 Extract gibt es noch die Möglichkeit der Einstellung der maximalen und mininalen Wörter der Termini. Die Qualität ist von ring“ (Noise) bis „Hoch“ (Silence) begrenzt. In Abbildung 2.11 wird die Standard-Füllwörterliste aus SDL MultiTerm 2014 Extract dargestellt.

[26] Quelle: Hilfe zu SDL MultiTerm 2014 Extract

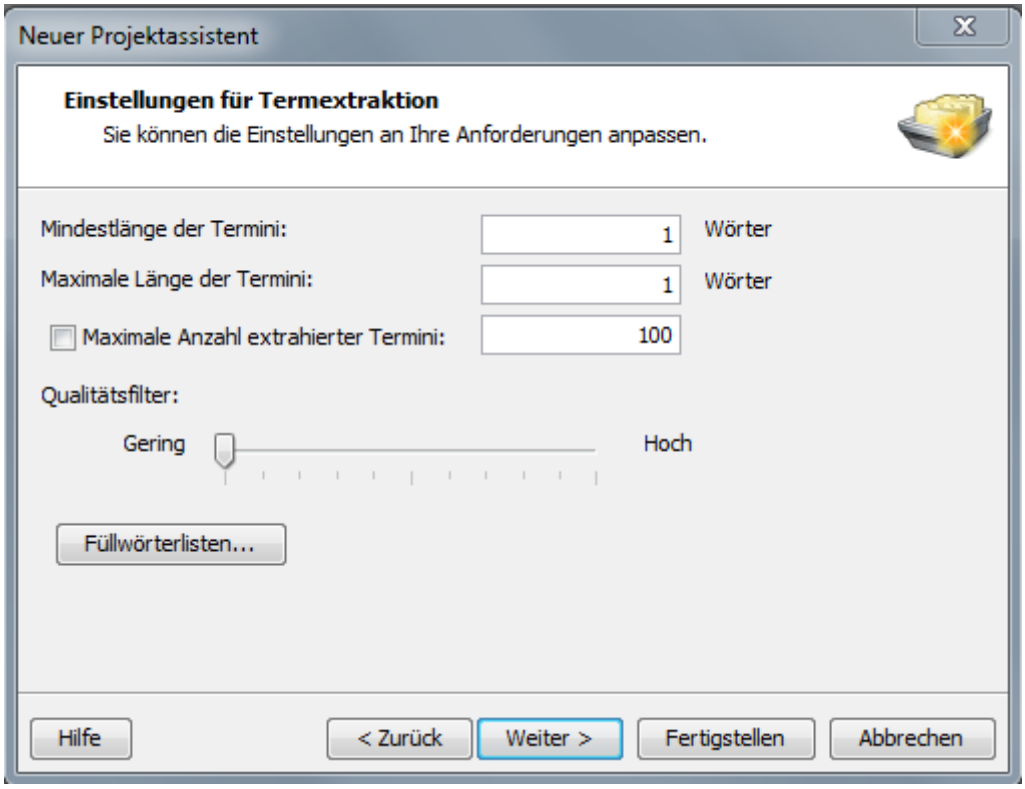


Abbildung 2.11 SDL MultiTerm 2014 Extract - Einstellungen für Termextraktion

Es gibt hier auch die Möglichkeit eine eigene Stoppwortliste hinzuzufügen (siehe Abbildung 2.12).

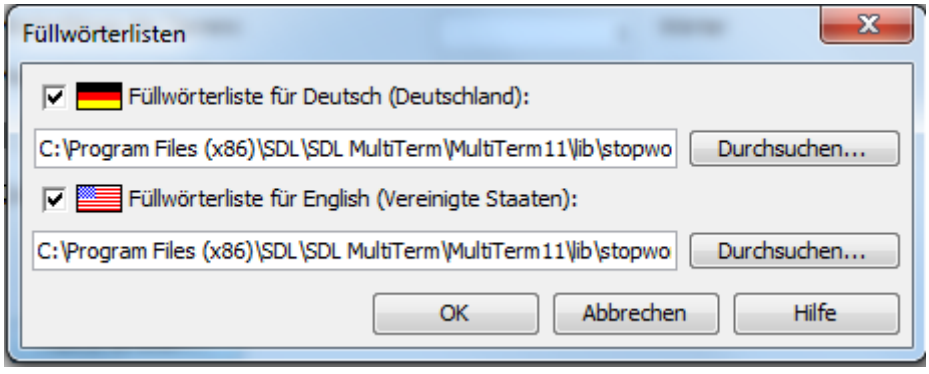


Abbildung 2.12 SDL MultiTerm 2014 Extract - Füllwörterlisten

Bei der zweisprachigen Termextraktion können ebenso die maximale Anzahl der Übersetzungen und die minimale Übersetzungshäufigkeit eingestellt werden (siehe Abbildung 2.13). Das Hinzufügen einer eigenen Stoppwortliste ist hier auch möglich.

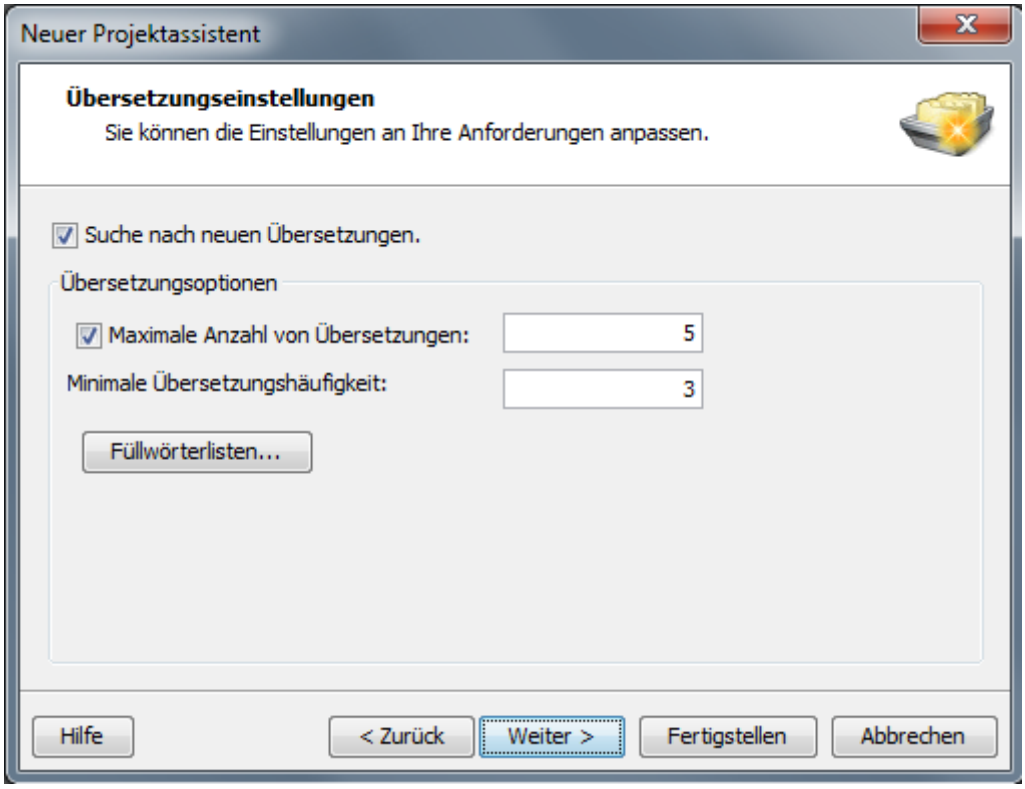


Abbildung 2.13SDL MultiTerm 2014 Extract - Übersetzungseinstellungen

Durch Klick auf die Schaltfläche „Fertigstellen“ startet der Extraktionsprozess. Die Extraktionsergebnisse werden in Form einer zweisprachigen Tabelle angezeigt. Bei weiterer Extraktion wird der Prozess mit der Funktion „Projekt“ > „Ausführen“ erledigt. Nach der Überprüfung der einzelnen Termkandidaten werden die Termini in dem Hauptfenster oder im Fenster linksunten bestätigt und bearbeitet. Die Termini werden durch Klick auf die Kontrollkästschen vor ihr bestätigt. Im Fenster linksunten können auch die Zusatzinformationen hinzugefügt werden. Leider werden nur die Synonyme in SDL MultiTerm 2014 Desktop erscheint. Durch Klick auf die Schaltfläche „Sätze generieren“ können der Kontext (Key word in context - KWIC) sowie die Quelle (unter dem Kontext mit Grau) angezeigt werden. Rechtsunten im Konkordanzfenster können die englischen Termini nach Markierung mit dem rechten Mausklick als Übersetzung hinzugefügt werden.

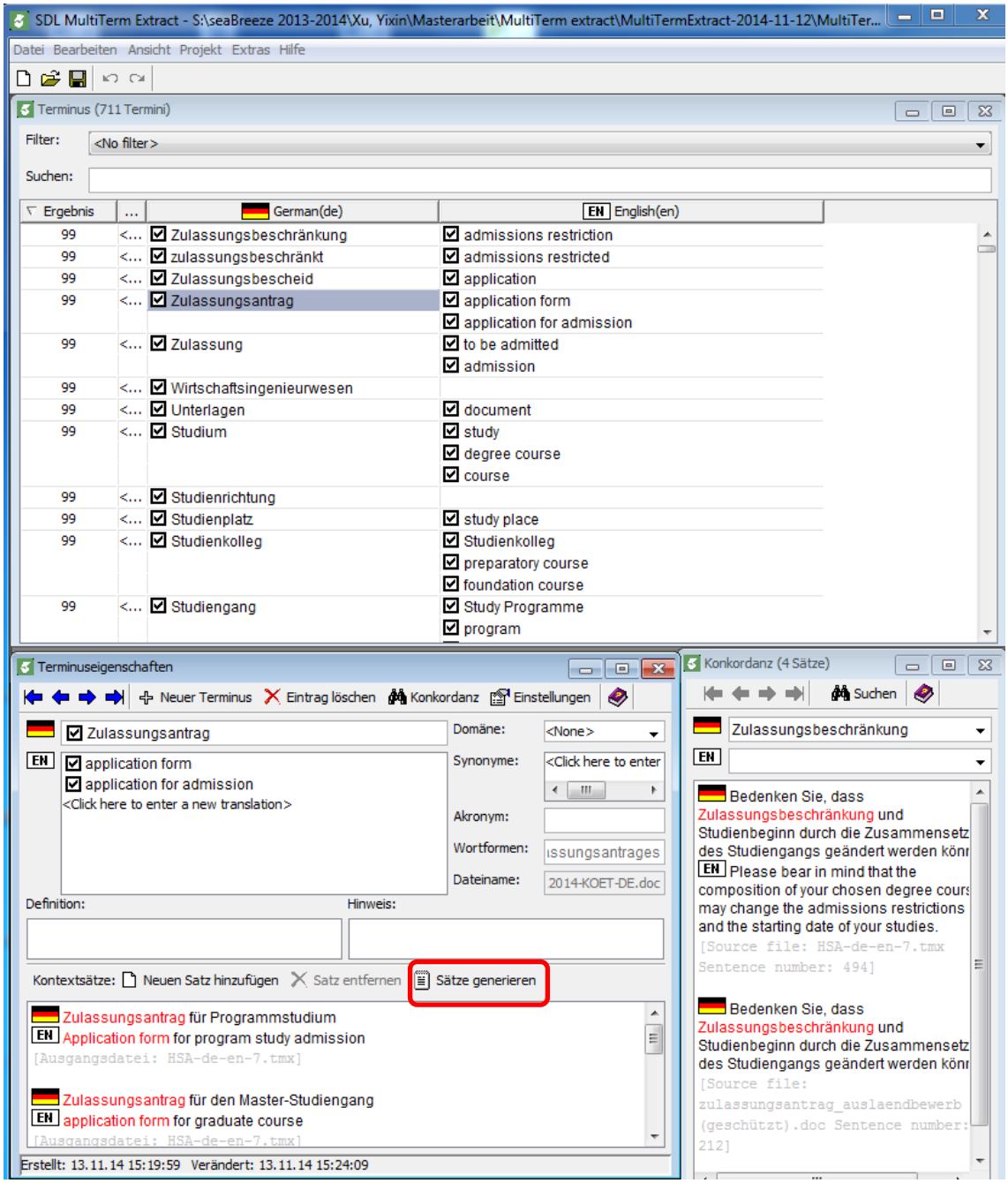


Abbildung 2.14 SDL MultiTerm 2014 Extract - Ansicht der Termextraktion

SDL MultiTerm 2014 Extract bietet nicht nur die automatische Extraktion, sondern auch eine manuelle Termextraktion. Mit der Funktion „Ansicht“ > „Textfenster“ wird zuerst ein Text ausgewählt. Dann wird ein Dialogfenster, wie in Abbildung 2.15 angezeigt, geöffnet. Durch Klick auf den neuen Termkandidaten mit der rechten Maustaste wird er als neuer Terminus hinzugefügt. Danach wird dieser Terminus mit rot markiert. Wenn eine automatische Termextraktion vorher durchgeführt wird, werden die extrahierten Termkandidaten auch rot makiert. Die Zahl der Seite der Datei wird im Dialogfenster unten links angezeigt.

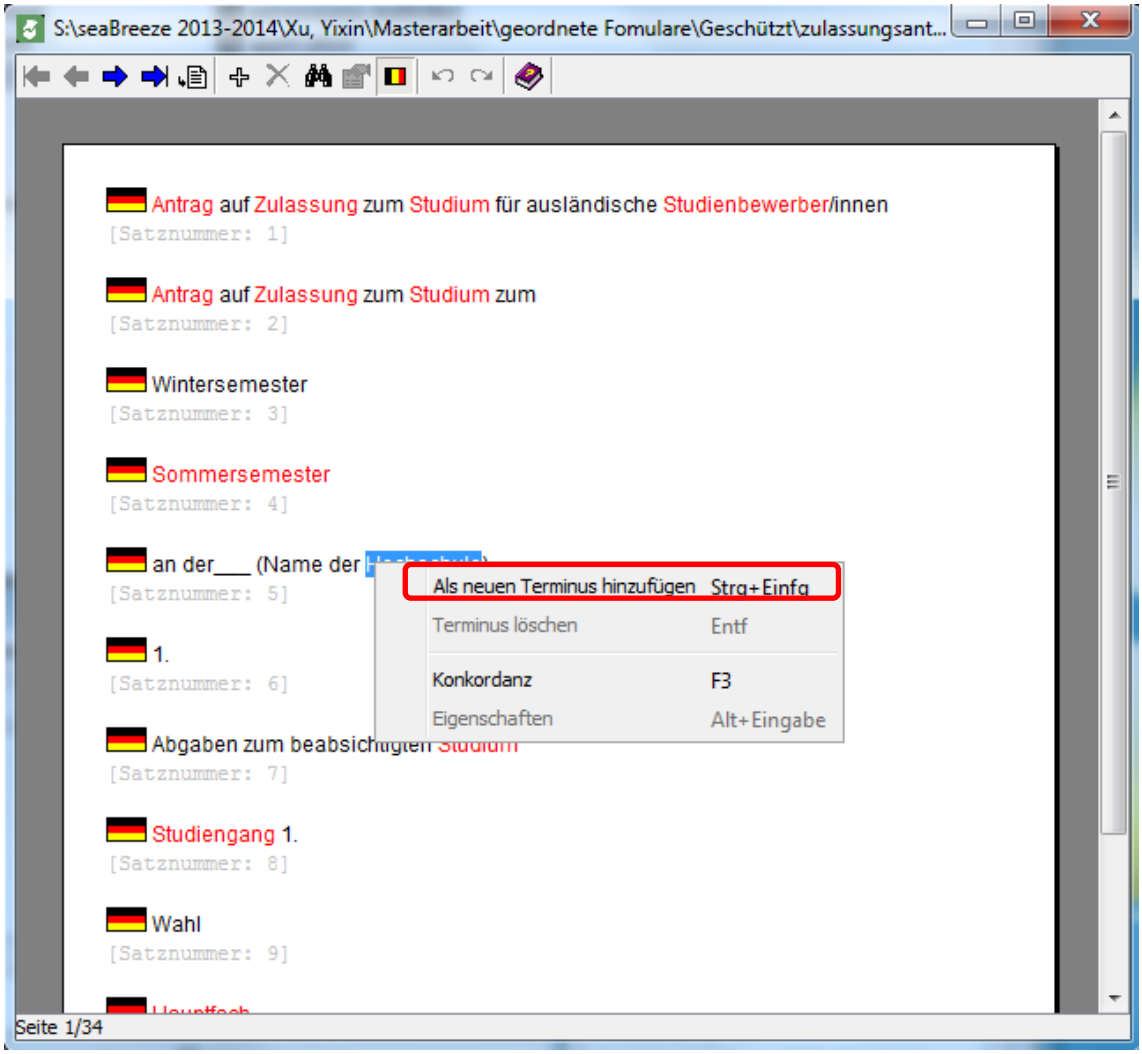


Abbildung 2.15 SDL MultiTerm 2014 Extract - manuelle Termextraktion

Nach dem Bestätigen der Termkandidaten können die extrahierten Termini direkt in der vorher hinzugefügten Projekttermbanken oder in MultiTerm XML bzw. in ein tabulatorgetrenntes TXT-Format exportiert werden (siehe Abbildung 2.16).

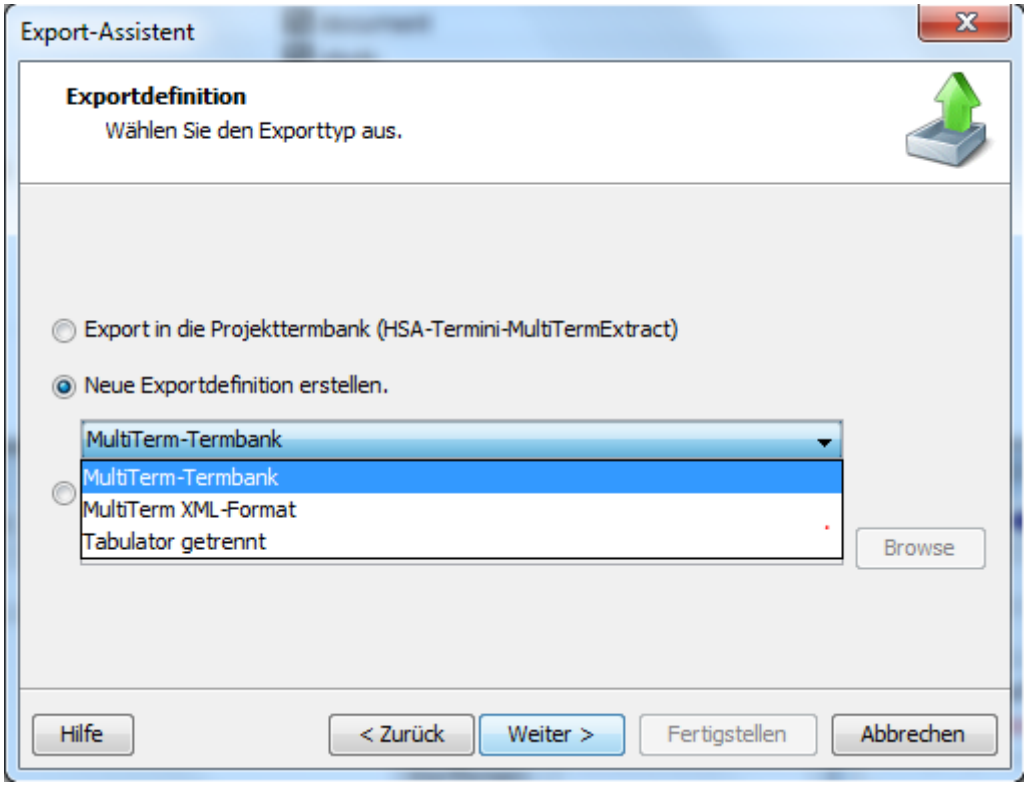


Abbildung 2.16 SDL MultiTerm 2014 Extract - Exportdefinition

Weiterhin kann ein Filter für die Exportdatei eingesetzt (siehe Abbildung 2.17) oder ein vorhandener Filter bearbeitet werden.

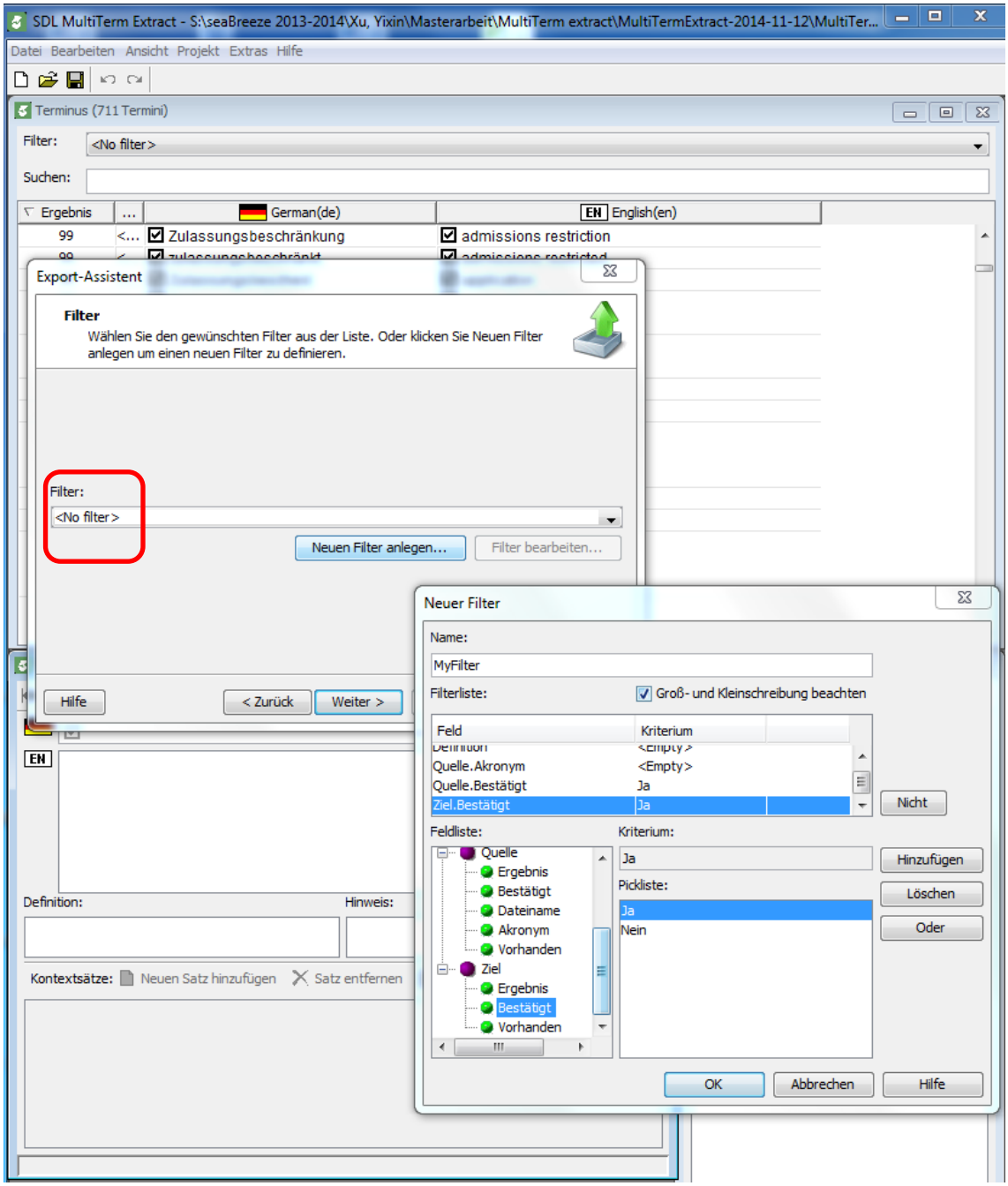


Abbildung 2.17 SDL MultiTerm 2014 Extract - Exportdefinition mit Filter

2.5.2 memoQ

Das Werkzeug memoQ unterstützt die Terminologieextraktion mit statistischen Verfahren. Die Projekteinstellungen in memoQ heißen „Sitzungen“, wo die akzeptierte Länge, die Häufigkeit der Kandidaten eingestellt werden können. Mit memoQ kann Termkandidaten aus den Ausgangsdokumenten, LiveDocs-Korpora oder Translation Memories extrahiert werden. Nach dem Hinzufügen von Ausgangsdateien werden die Termkandidaten durch Klick auf die Option

„Begriffe extrahieren...“ oder „Kandidaten extrahieren“ [27] im Menü „Vorgänge“ aufgelistet.

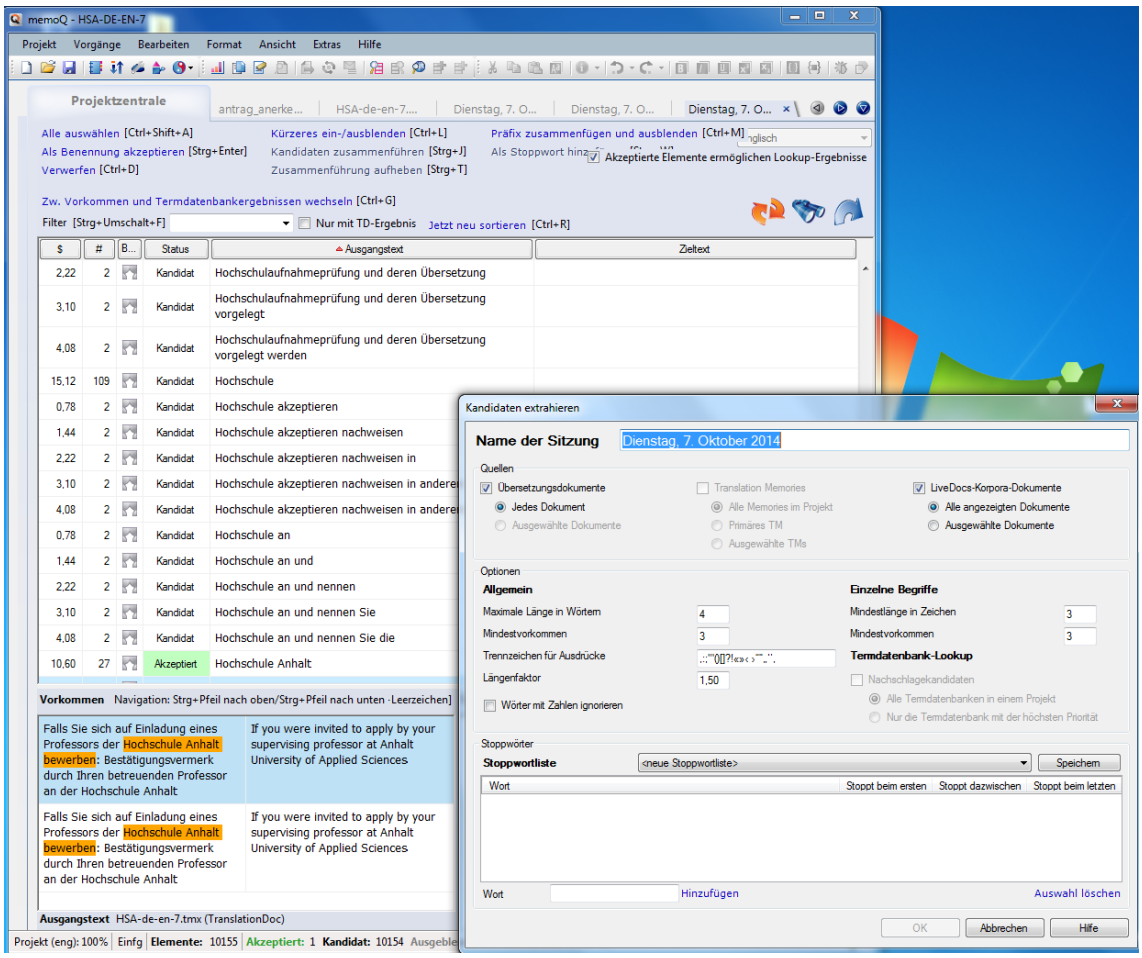


Abbildung 2.18 memoQ Extraktionsoberfläche

Im Vergleich zu SDL MultiTerm 2014 Extract hat memoQ Vorteile und auch viele Nachteile. Die Gemeinsamkeiten sowie Unterschiede zwischen den beiden Werkzeugen werden in der folgenden Tabelle aufgelistet.

Vergleich zwischen SDL SDL MultiTerm 2014 Extract und memoQ		
Funktionen	**SDL MultiTerm 2014 Extract**	**memoQ**
Dateiformate	Wenig (dieses Projekt txt,word 2003-2007, tmx)	Umfangreich (dies Projekt pdf,word,tmx)
Technik	statistisch	statistisch

[27] Bei der ersten Terminologieextraktion im aktuellen Projekt wird das Dialogfeld „Kandidaten extrahieren“ angezeigt.

Silence und Noise	Änderungsmöglichkeit im Prozess	Einmalige Einstellung
Stoppwortliste	Liste aus MultiTerm Extract, bzw. die Möglichkeit zum Hochladen einer neuen Liste.	Selbstes Hinzufügen bei der Einstellung
Zusatzinformationen/ Zusatzfunktionen	Synonym, Akronym, Definition, Kontext, Bemerkung usw (aber nicht alle funktionieren)	Vorschau
Wiederholungen	Meldung beim Bearbeiten	Keine Meldung beim Bearbeiten
Nachbearbeitungen	Verloren bei großer Änderung den Kontext	Anzeige der Originalbezeichnung
Eintrag von Synonym	ja	Nein
Sprache	Monolingual, bilingual	Monolingual
Export	Mtx, mtb	

3 Vorgehensweise im Projekt

In „Terminologiearbeit-Best Practices[28]“ werden die Terminologieprozesse beim Unternehmen so beschrieben:

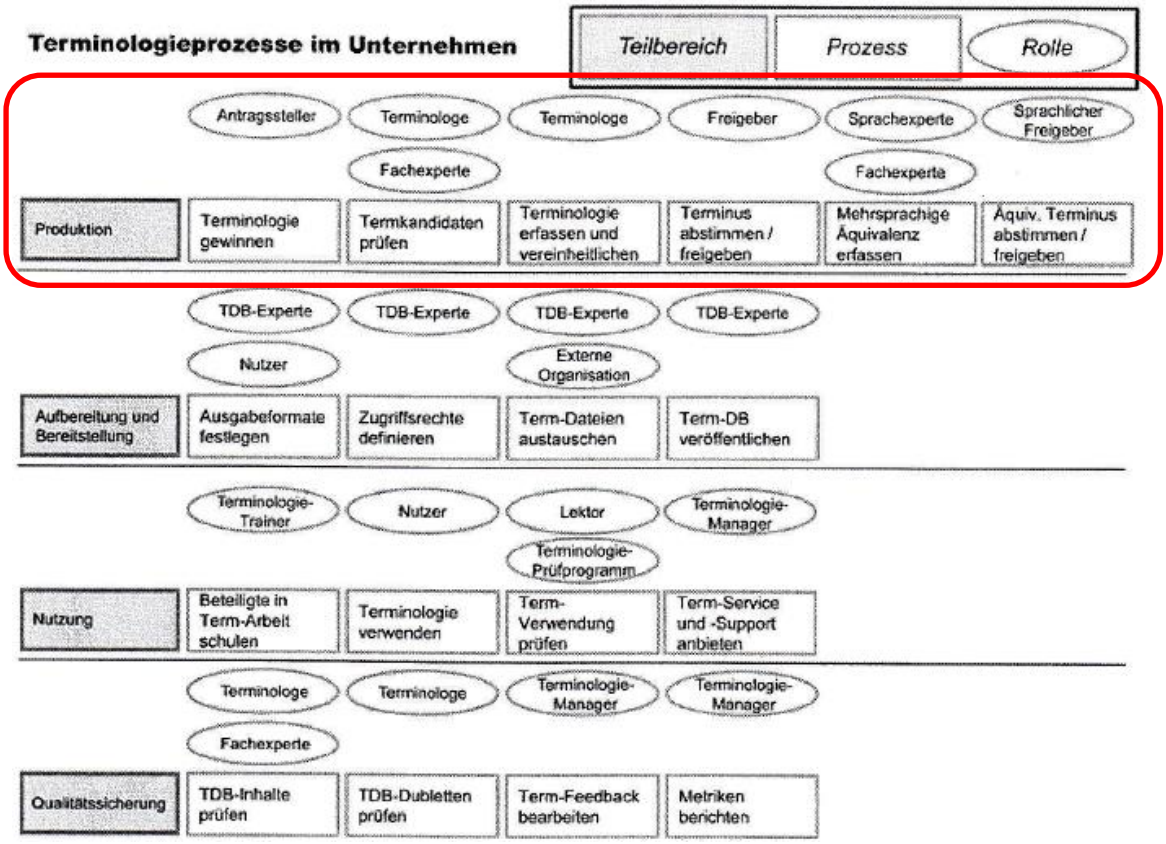


Abbildung 3.1 Terminologieprozesse im Unternehmen [29]

In dieser Arbeit wird der Prozess im Teilbereich „Produktion“ verwendet, nämlich Terminologie gewinnen, Termkandidaten prüfen, Terminologie erfassen und vereinheitlichen, Termini abstimmen/ freigeben, mehrsprachige Äquivalenz erfassen und äquivalente Termini abstimmen/ freigeben. Dieser Schritt wird angenommen. Aber die Reihenfolge dieses Prozesses wird bei dieser Arbeit geändert. In dieser Arbeit werden zuerst die Ausgangsmaterialien vorbereitet und analysiert. Dann wird die bilinguale Zuordnung von Äquivalenzen und monolin-

[28] [Arndt 2014]

[29] Quelle: M5-12 in Terminologiearbeit – Best Practices

gualer Extraktion von Termkandidaten ausgeführt. Das Ergebnis bzw. der Inhalt der Termbank wird danach von der Projektkoordinatorin, Frau Prof. Dr. Uta Seewald-Heeg, und Herrn Dr. Horst Seiler überprüft. Die bestätigten Termini bzw. seine Äquivalenzen werden in eine vordefinierte Datenbank importiert und dann freigegeben.

3.1 Analyse der Ausgangsmaterialien und Bestimmung des Datenvolumens

Der erste Schritt der Terminologieextraktion ist die Vorbereitung und Analyse von vorhandenen Ausgangsmaterialien, z. B. Dateiformaten. Dabei werden die Texte analysiert, um Wörter und Phrasen zu finden und herauszufiltern. Mithilfe von SDL Trados Studio 2014 werden die Ausgangsmaterialien analysiert und der Umfang dieses Projektes wird bestimmt.

Vor der Analyse sind einige Voraussetzungen notwendig. Die erste Voraussetzung ist, dass alle Ausgangsdateien nach den Anforderungen in verschiedenen Ordnern zugeordnet werden müssen. Da verschiedene Extraktionswerkzeuge verschiedene Dateiformate unterstützen, müssen die Dateiformate vor der Extraktion bzw. vor der Analyse vorbereitet werden. SDL MultiTerm 2014 Extract unterstützt PDF-Dateien nicht, deswegen müssen die PDF-Dateien in Microsoft Word 97-2003 oder TXT-Dateien konvertiert werden. Mithilfe der Software Adobe Acrobat Pro können die PDF-Dateien als Microsoft Word 97-2003 und MS-Word gespeichert werden. Sieben PDF-Dateien enthalten XFA-Formulare, so dass die Antragsteller die Inhalte in die Formulare eintragen können. Solche PDF-Dateien können nicht als Word-Dateien sondern als TXT-Dateien gespeichert werden.

Nach der Analyse werden die Informationen in folgender Tabelle gesammelt. Zuerst gibt es einen Überblick über den Umfang des Projektes. Und dazu werden die allgemeinen Informationen des Projektes zusammengefasst. Es gibt insgesamt 47 PDF-Dateien mit 22789 Wörtern. Dazu sind drei PDF-Dateien geschützt, die nicht bearbeitet, kopiert oder in andere Dateiformate konvertiert werden können. Das heißt es gibt 6958 Wörter, die manuell in Word-Dateien

eingegeben müssen. Oder die Termkandidaten werden in gedrucktem Papier von Hand markiert, und dann in MultiTerm einzel eingetragen.

Für allgemeine Informationen						
	Datei	Deutsch	Englisch	Deutsch und Englisch	Segmente	Wörter
Bernburg	10	9		1	800	3966
Dessau	16	12	2	2	884	3635
Köthen	18	15		3	1767	8330
Geschützt (Köthen)	3	1		1[30]	696	6858
Alle	47	37	2	7	4147	22789

Nach den Inhalten und den Behandlungen der Dateiformate werden drei Methoden, mit MS-Excel bzw. MS-Word, mit SDL MultiTerm 2014 Extract und mit SDL Trados Studio 2014, zur Extraktion bestimmt. Dazu werden folgende Informationen gesammelt, um jeweils ein genaues Konzept oder Verfahren festzulegen.

Bei der Textsorte der Ausgangsmaterialien handelt es sich um Anträge. Darin gibt es viele tabellarische Inhalte. Einige Dateien enthalten wenig Fließtext. Aus diesem Grund wird folgende Tabelle erstellt. Die Ausgangsdateien enthalten vier Dateien, die einen identischen Inhalte wie die anderen vier Dateien haben. Die Termkandidaten in den mit wenigen Fließtexten erfassten Dateien werden nach dem Durchlesen manuell in Excel-Datei eingegeben. Für die Terminologieextraktion aus den mit mehreren Fließtexten geschriebenen Dateien wird zuerst eine Wortliste mit Hilfe von MS-Word erzeugt und dann in einer Excel-Datei bereinigt und bearbeitet. Insgesamt sind 42[31] PDF-Dateien mit 19751 Wörtern zu bearbeiten.

[30] Es gibt 2 deutschen Dateien und 1 englische Datei. Davon ist 1 deutsche Datei die Übersetzung von der englischen Datei. Die beiden Dateien werden in einer Datei zusammengefürt.

[31] Die drei einsprachigen geschützten Dateien werden jeweils zu DE-EN mit mehreren Fließtexten und DE mit mehreren Fließentexten zugeordnet.

Mit einem einfachen Verfahren			
	Dateien	Segmente	Wörter
Inhaltlich identische Dateien	4	532	2805
DE mit wenigen Fließtexten	25	1138	3963
DE mit mehreren Fließtexten	9	1182	5145
EN mit wenigen Fließtexten	2	55	218
EN mit mehreren Fließtexten	-	-	-
DE-EN mit wenigen Fließtexten	2	96	411
DE-EN mit mehreren Fließtexten	4	1118	10014
Alle außer inhaltlich identischen Dateien	42	3589	19751

Der Schwerpunkt bei der automatischen Terminologieextraktion in diesem Projekt ist die zweisprachige Extraktion. Bei der automatischen Terminologieextraktion sind die englischen Termkandidaten einfach herauszufinden, wenn die entsprechend deutschen Termkandidaten hervorgehoben werden. Aus diesem Grund werden die Informationen der englischen Dateien, die aus den gemischten Dateien entnommen werden, nicht erfasst. Die folgende Tabelle zeigt, dass nur 1138 Segemente mit 6321 Wörtern nach der automatischen Extraktion durchgelesen werden.

Automatische Extraktion oder mit alternativen Methoden							
	Dateien	neue Segmente	neue Wörter	wiederholte Segmente	wiederholte Wörter	alle Segemente	alle Wörter
DE	37	1053	5982	1572	4208	2625	10190
DE in gemischten Dateien	7	49	180	941	5437	990	5617
EN	2	36	159	19	59	55	218
Gesamt	46	1138	6321	2532	9704	3670	16025

3.2 Extraktion

Es gibt vielfältige Methoden bzw. Werkzeuge zur Terminologieextraktion. Wegen der Beschränkung der Arbeitszeit sowie der verfügbaren Werkzeuge werden nur zwei Verfahren, ein einfaches Verfahren mit MS-Excel und MS-Word und ein statistisches Verfahren mit dem Werkzeug SDL MultiTerm 2014 Extract, in diesem Projekt verwendet. Die Terminologieextraktion mit dem einfachen Verfahen hat bessere Qualität, während die Extrakton mit SDL MultiTerm 2014 Extract sehr schnell ist. Um eine bessere Extraktionsqualität zu garantieren, steht die Terminologieextraktion mit dem einfachen Verfahren im Mittelpunkt.

3.2.1 Extraktion mit einem einfachen Verfahren

Vor der Extraktion muss zuerst eine optimale Struktur für die Termbank in MS-Excel festgelegt werden. Mit dieser Struktur können die extrahierten Termkandidaten in MultiTerm problemlos importiert werden. Die Struktur der beschreibenden Felder in Excel wird in Abbildung 3.2 mit einigen Beispielen angezeigt. Die Bezeichnungen in der Überschriftzeile wie Definitionen, Kontextbeispiele usw. sollen mit den Bezeichnungen der beschreibenden Felder identisch sein. Alle Synonyme eines Begriffs werden unter gleicher Eintragsnummer bearbeitet.

	A	C	D	E	F	L	M	N	O	P	Q	R	S	T	X	Y	Z	AA	AB
1	Begr	DE	Defin	Defin	Zugr	Quel	Zugr	Kont	Quel	Zugr	EN	Defi	Defi	Zugri	Quel	Zugr	Kont	Quel	Zugr
2	1	1. Prüfer				antrag_abschlussarbeit.p					initial examiner				http:	2014-11-03-11:00			
3	1	Erstprüfer				90-DMP_AntragZulassungMas					first examiner				http:	2014-11-03-11:00			
4	2	2. Prüfer				antrag_abschlussarbeit.p					second examiner				http:	2014-11-03-11:02			
5	2	Zweitprüfer				90-DMP_AntragZulassungMas					second moderator				http:	2014-11-03-11:03			
6	3	3. Prüfer				zulassung_fachpruefung.p					third-party auditor				http:	2014-11-03-11:06			
7	3	Drittprüfer				http:	2014-11-03-11:04				third examiner				http:	2014-11-03-11:09			
8	4	Abdruck				BA_Berufs	Im Versicherung				copy				http:	2014-11-03-11:12			
9	5	Abendgymnasium	Das Al	http:	2014-	Zulant_BA2014-KOET-DE.pdf					evening college				Zulant_MA2014-BBG_01.PDF				
10	5	Abendgymnasium									evening gymnasium				http:	2014-11-10-11:06			
11	5	Abendgymnasium									Abendgymnasium	An A	http	2014-	http:	2014-11-10-11:06			
12	6	Abgabe				90-DMP_AufgabenstellungMa					submission				90-Master_Aufgabenstellu				
13	7	Abgabedatum				antrag_abschlussarbeit.p					handover date				http:	2014-11-03-11:15			
14	7	Abgabetermin				BachelorAr	Neuer Abgabeter				deadline				http:	2014-11-03-11:16			
15	7	Abgabetermin									closing date				http:	2014-11-03-11:17			
16	8	Abgabeort				AntragZulassungBachelorDi					giving down place				http:	2014-11-03-11:18			
17	9	abgeschlossene Berufsausbildung				Zulant_FS2	Abgeschlossene				professional training to date and				Zulant_MA2014-BBG_01.PDF				
18	9	abgeschlossene Berufsausbildung									completed vocational training				http:	2014-11-26-11:18			
19	10	abgeschlossenes Hochschulstudium				Zulant_MA2	Dieser Antrag l				achieved university qualification				Zulant_MA2	this applicati			
20	11	Abitur	Das Ab	http:	2014-	zulassungs	In bundesweiter				Abitur	Abit	http	2014-	applicati	EU foreigners			
21	11	Abitur									grammar school leaving certificate				applicati	EU foreigners			
22	12	Ablehnung				zulassungs	Eine nicht wahr				reject				applicati	An incorrect a			
23	12	Ablehnung									rejection				DAAD-Wörterbuch				
24	13	Ableistung				A-Beurlaub	die Vorschrifte				completion				applicati	the regulation			
25	14	Abschluss				zulassungs	Bitte geben Sie				award of the school leaving certif				applicati	Please provide			
26	14	Schulabschluss	Seiner	http:	2014-	zulassungsantrag_auslaen					end of the schooling				application_foreignstude				

Abbildung 3.2 Struktur der Termbank in Excel

Damit die zweisprachigen Texte parallel angezeigt werden, wird hier die Funktion „Text in Tabelle umwandeln“ in einer Word-Datei verwendet. Die Ausgangstexte und Zieltexte werden durch Absatzmarken getrennt. Zwischen den Ausgangstexten oder den Zieltexten muss keine Absatzmarke gesetzt werden. Dann kann die Form der Anzeige von Texten problemlos in MS-Word mit der Funktion „Hinzu > Tabelle > Text in Tabelle umwandeln“ geändert werden. Die Texte werden in zwei Spalten durch Absätze getrennt. (siehe Abbildung 3.3)

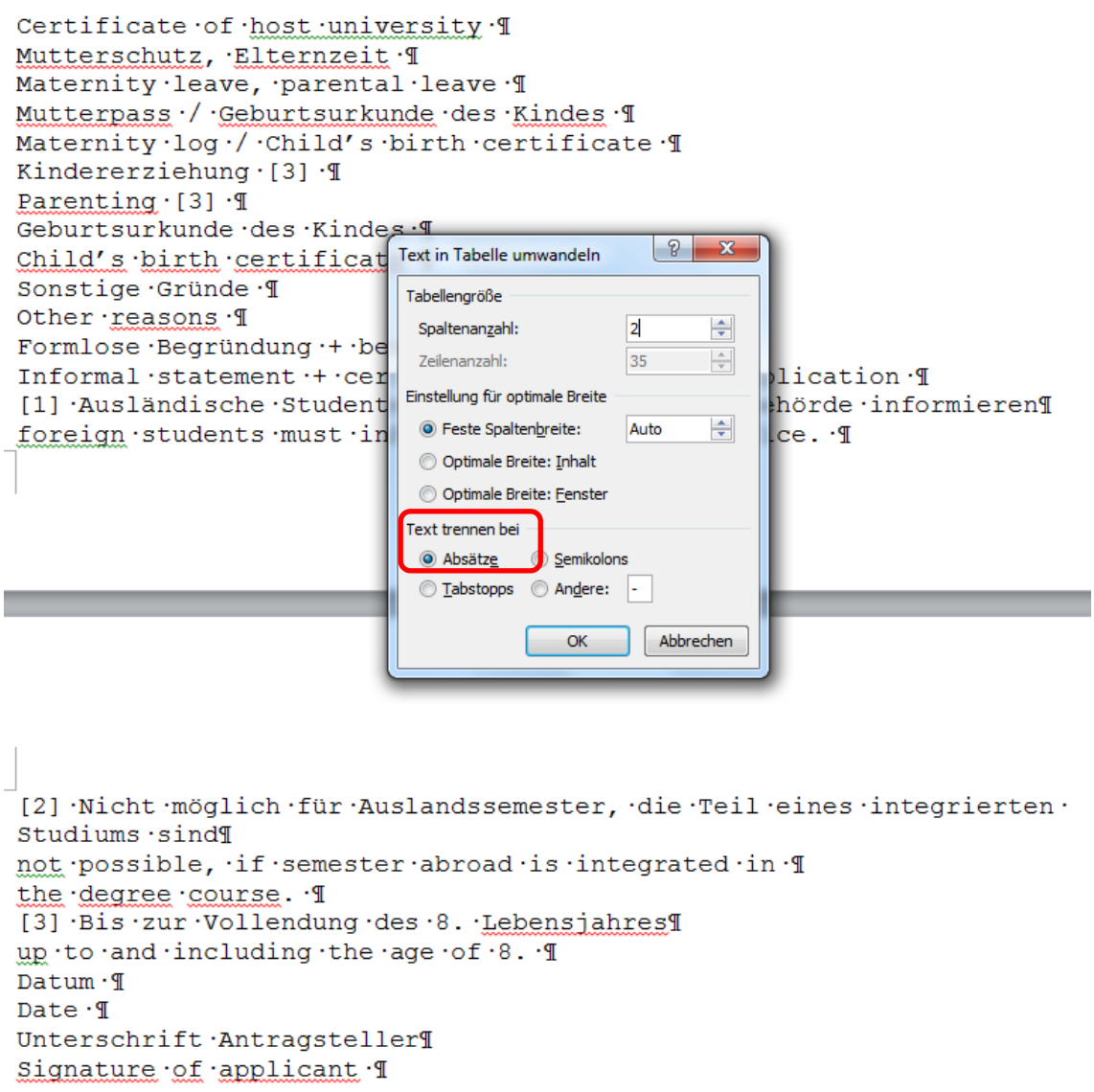
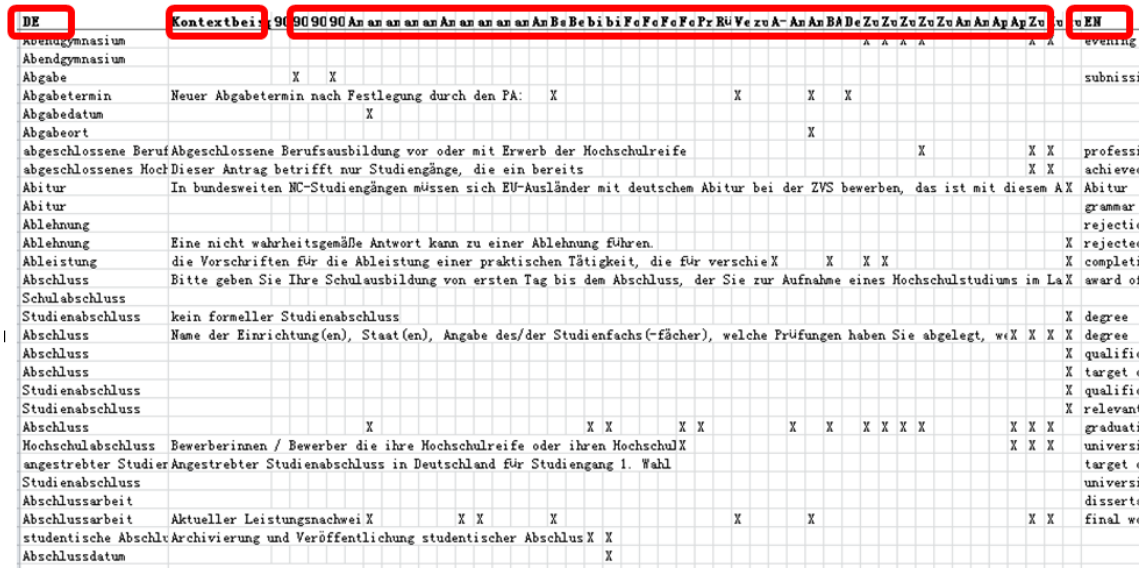


Abbildung 3.3 Text in Tabelle umwandeln

Bei der Termextraktion werden zuerst die gemischten Dateien, die wenige Fließtexte haben, manuell in eine Excel-Datei extrahiert. Wenn eine zweisprachige Dateie relativ viele Fließtexte enthält, wird sie durch die in Kapitel 2.3.1 vorgestellte Methode, mit der Funktion „suchen und ersetzen“ in MS-Word, in

Liste überführt. Die Quellen der Termini, die Kontextbeispiele sowie deren Quellen werden ebenfalls in der gleichen Zeile hintereinander eingefügt. (siehe Abbildung 3.4)



DE	Kontextbei	EN
Abendgymnasium		evening
Abendgymnasium		
Abgabe		subnissi
Abgabetermin	Neuer Abgabetermin nach Festlegung durch den PA:	
Abgabedatum		
Abgabeort		
abgeschlossene Beruf	Abgeschlossene Berufsausbildung vor oder mit Erwerb der Hochschulreife	professi
abgeschlossenes Hoch	Dieser Antrag betrifft nur Studiengänge, die ein bereits	achieve
Abitur	In bundesweiten NC-Studiengängen müssen sich EU-Ausländer mit deutschem Abitur bei der ZVS bewerben, das ist mit diesem A	Abitur
Abitur		grammar
Ablehnung		rejecti
Ablehnung	Eine nicht wahrheitsgemäße Antwort kann zu einer Ablehnung führen.	rejecte
Ableistung	die Vorschriften für die Ableistung einer praktischen Tätigkeit, die für verschie	completi
Abschluss	Bitte geben Sie Ihre Schulausbildung von ersten Tag bis dem Abschluss, der Sie zur Aufnahme eines Hochschulstudiums im La	award o
Schulabschluss		
Studienabschluss	kein formeller Studienabschluss	degree
Abschluss	Name der Einrichtung(en), Staat(en), Angabe des/der Studienfachs(-fächer), welche Prüfungen haben Sie abgelegt, w	degree
Abschluss		qualifi
Abschluss		target
Studienabschluss		qualifi
Studienabschluss		relevan
Abschluss		graduati
Hochschulabschluss	Bewerberinnen / Bewerber die ihre Hochschulreife oder ihren Hochschul	universi
angestrebter Studier	Angestrebter Studienabschluss in Deutschland für Studiengang 1. Wahl	target
Studienabschluss		universi
Abschlussarbeit		dissert
Abschlussarbeit	Aktueller Leistungsnachwei	final w
studentische Abschlu	Archivierung und Veröffentlichung studentischer Abschlus	
Abschlussdatum		

Abbildung 3.4 Kontextbeispiele und Quelle in Excel

Für den Import in eine begriffssortierte Datenbank sind hier die Kontextbeispiele besonders wichtig, so dass andere Bedeutungen einer Benennung nicht erfasst werden. Beim Auslassen eines Wortbestandteils sind die Kontextbeispiele auch ein guter Hinweis zur Überprüfung, ob die verkürzte Benennung ein Synonym der originalen Benennung ist. Mit dem selben Beispiel wie in Kapitel 2.2.3 im Satz „Bitte geben Sie Ihre Schulausbildung vom ersten Tag bis zum Abschluss, der Sie zur Aufnahme eines Hochschulstudiums im Land der Ausstellung berechtigt, an.“ gibt es das Wort „Abschluss“. Hier bedeutet es „Schulabschluss“ aber nicht „Studienabschluss“ oder „Hochschulabschluss“. In der Datenbank wird „Abschluss“ als Synonym von „Schulabschluss“ bezeichnet. Gleichzeitig werden die Synonyme und Abkürzungen ausgewählt und sortiert, so müssen diese Termini bei der späteren Bearbeitung anderer Termkandidaten nicht nochmal bearbeitet werden (siehe Abbildung 3.5).

1	1. Prüfer		18	Studiengangswechsel
1	Erstprüfer		18	Studiengangwechsel
2	2. Prüfer		18	Wechsel des Studiengang
2	Zweitprüfer		19	Wintersemester
3	Abschlussarbeit		19	WiSemester
3	studentischer Abschlussarbeit		19	WS
4	Abgabedatum		20	wissenschaftlicher Betreuer
4	Abgabetermin		20	Betreuer
5	akad.Grad		21	Bachelorstudiengang
5	akademisch Grad		21	Bachelor-Studiengang
6	Aufgebenstellung		22	Studiengang
6	Thema		22	Stg.
7	Bescheid		23	Wiederholungsprüfung
7	Mitteilung		23	Nachprüfung
8	Beschreibung der Arbeit		24	erstmalige Immatrikulation
8	Inhaltsangabe		24	Erstimmatrikulation
9	Fach		25	Eintritts-Fachsemester
9	Modul		25	Eintrittssemester
10	Hochschule Anhalt		26	Berufliche Ausbildung
10	Hochschule Anhalt (FH)		26	Berufsausbildung
10	HOCHSCHULE ANHALT (FH)		27	berufliche Tätigkeit
10	HSA		27	Berufstätigkeit
11	Masterarbeit		28	Fachgebundene Hochschulreife
11	Masterthesis		28	Fachhochschulreife
12	Masterstudiengang		29	NC
12	MA-Studiengang		29	Eignungsfeststellung
13	Matr.-Nr.		30	Duale-Studiengang
13	Matrikel-Nr.		30	duale Studiengang
13	Matrikelnummer		31	Zweitstudium
14	Meldung		31	zweites Studium
14	Zulassung		32	NC-Studiengang
15	Prüfungsamt		32	zulassungsbschränkter Studieng
15	Prüfungsausschuss		33	Kfz.-Kennz.
15	Prüfungsausschuß		33	Kfz-Kennz.
16	Scheine		33	Kfz-Kennzeichen
16	Vorleistung		34	Eingangsstempel
17	Sommersemester		34	Vermerke HSA
17	SoSemester		35	Internat. Business (IBS)-englisch
17	SS		35	Internat. Business (IBS)- engl. Zw

7DE-EN | EN | DE | **Synonym** | Abk. | alle Extraktion | Tabelle2

Abbildung 3.5 Tabellenblatt in Excel

Die einsprachigen deutschen und englischen Dateien werden in gleicher Art und Weise extrahiert. Dann werden alle Kandidaten in den vorher erstellten Synonym- und Abkürzung-Tabellen nach dem Begriff sortiert. Danach wird die Tabelle nach dem Dateinamen sortiert. Und das Symbol „X“ wird durch die Da-

teinamen ersetzt. Alle Dateinamen eines Terms werden mithilfe von Visual Basic[32] in Excel in einem Feld zusammengesammelt. Für die Ergänzung der Termbank werden die anderen Übersetzungen und Quellen sowie die Zusatzinformationen hinzugefügt. Da es immer mehr Kooperationen mit chinesischen Hochschulen gibt, werden auch die entsprechenden chinesischen Termini mit Hilfe des „Deutsch-Chinesischen Universitätswörterbuch“ erstellt.

Einige Wortkombinationen wie „kleinste Studieneinheit“, „1. Prüfer“, „Beschreibung der Arbeit“ werden auch als Termkandidaten extrahiert, die bei der automatischen Extraktion nicht extrahiert wurden.

In diesem Projekt werden die Prozesse „Übersetzung“ und „Erfassung zusätzlicher Daten“ in Excel durchgeführt. Schließlich werden die angeforderten Felder in einer bestimmten Reihenfolge hinzugefügt. Die Reihenfolge des Überschrifttitels ist wie folgt angeordnet.

- Begriffsnummer
- Sachgebiet
- DE
- Definition
- Definitionsquelle
- Zugriffsdatum
- Kommentar
- Wortart
- Genus
- Termtyp
- Status-DE
- Quelle
- Zugriffsdatum
- Kontextbeispiele
- Quelle Kontext
- Zugriffsdatum
- EN

[32] Unter der Menüleiste „Entwicklungstool“

- Definition
- Definitionsquelle
- Zugriffsdatum
- Kommentar
- Wortart
- Status
- Quelle
- Zugriffsdatum
- Kontextbeispiele
- Quelle Kontext
- Zugriffsdatum

3.2.2 Extraktion mit einem statistischen Verfahrensprogramm

In dieser Arbeit wird das Programm SDL MultiTerm 2014 Extract verwendet, das mit einem statistischen Verfahren arbeitet. Vor dem Extrahieren wird zuerst die Konvertierung von Daten vorbereitet. Die Ausgangstexte und Zieltexte aus den gemischten Dateien werden in MS-Word getrennt und gespeichert. Dann werden die Ausgangstexte mit entsprechenden Zieltexten in Trados Studio aligniert (siehe Abbildung 3.6). Die Alignments werden in ein vorher erstelltes leeres Translation Memory importiert und in TMX-Format exportiert. Die Qualität des Alignments ist schlecht, wenn die Ausgangstexte in mehreren Spalten und die einzelnen Punkt in mehreren Zeilen geschrieben werden (siehe Abbildung 3.7).

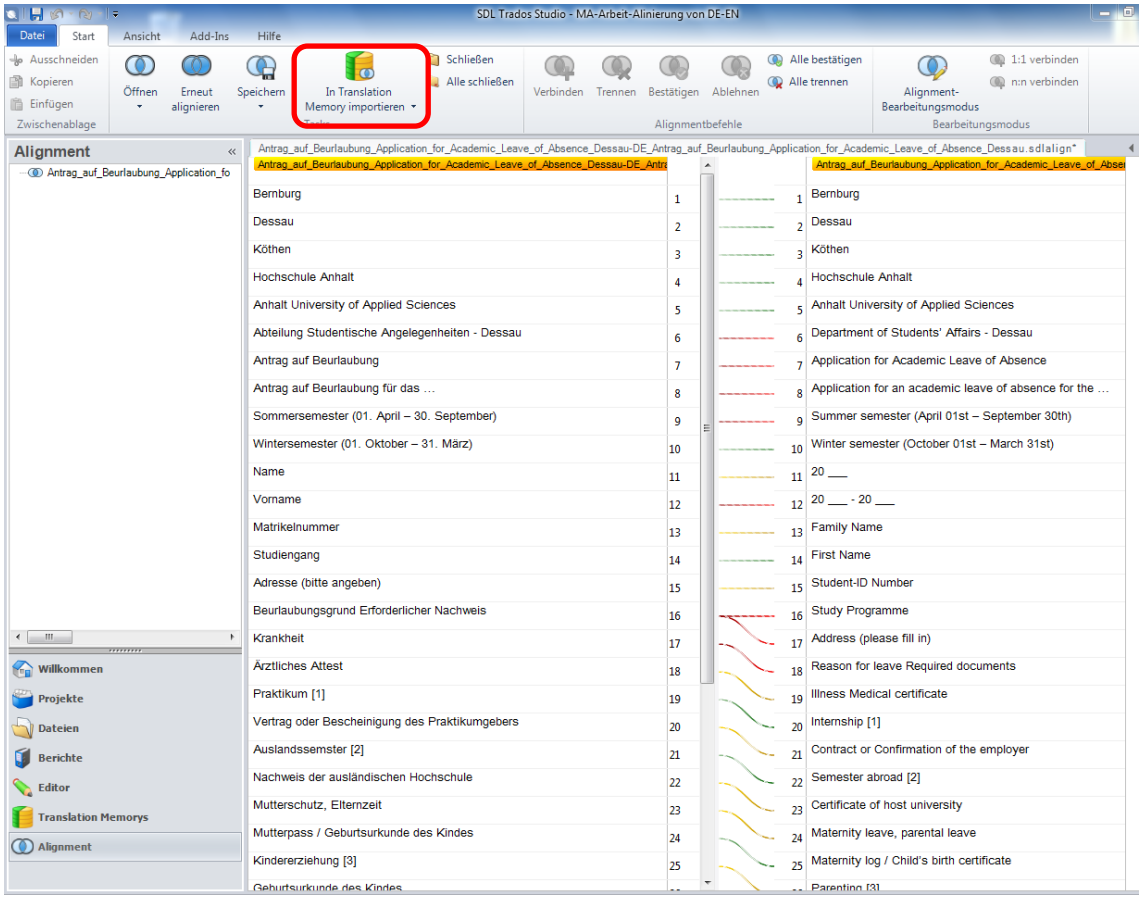


Abbildung 3.6 SDL Trados Studio 2014 - Alignment

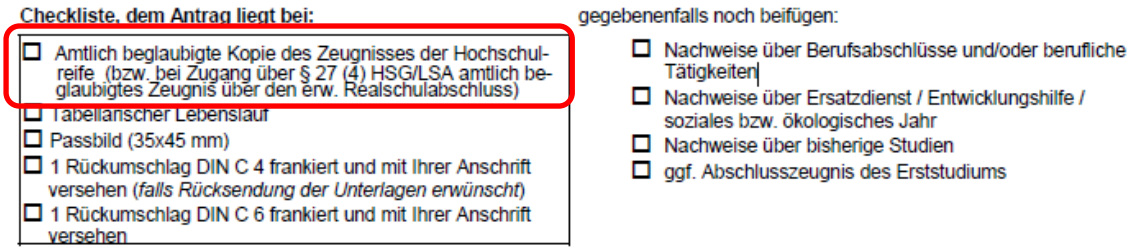


Abbildung 3.7 PDF-Datei - Problem beim Alignment

Wie in Kapitel 2.4.1 vorgestellt, wird zuerst ein Projekt erstellt. Dann werden die Termkandidaten extrahiert und bestätigt. Am Ende wird die Termbank in MultiTerm XML oder eine mit Tabulator getrennte Text-Datei exportiert. Nach der Analyse gibt es viele wiederholte Segmente und Wörter zwischen den einsprachigen und zweisprachigen Dateien. Um die wiederholten Wörter oder Kollokationen bei zweisprachiger Termextraktion nicht mehr anzuzeigen, wird hier die „zweisprachige Termextraktion“ ausgewählt, ohwohl es noch einsprachige Dateien gibt. (siehe Abbildung 3.8). Die Dateiformate sind bei SDL MultiTerm 2014

Extract auf das TXT-Format oder MS-Word 97-2003 beschränkt. [siehe Kapitel 2.4.1 unterstützte Dateiformate]

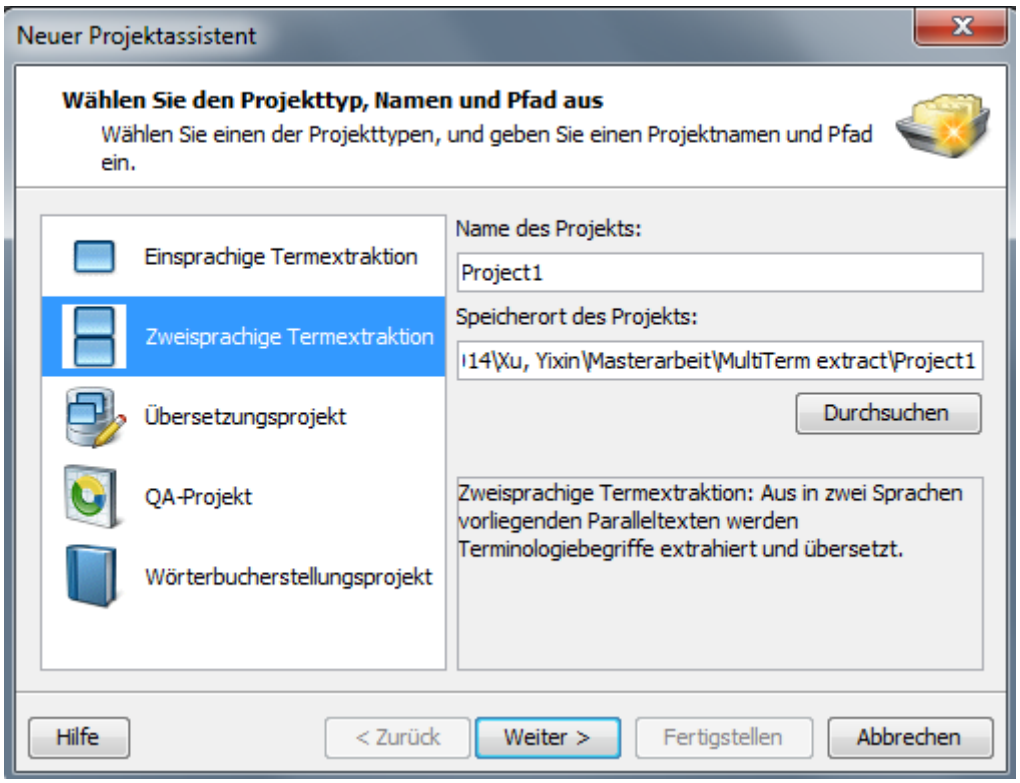


Abbildung 3.8 SDL MultiTerm 2014 Extract - Projekttyp als Zweisprachige Termextraktion auswählen

Die Mindestlänge der Termini wird bei der Extraktion öfters eingestellt. Wie Abbildung 3.9 angezeigt wird „eins“ als die Maximale Länge der Termini eingesetzt. Der Qualitätsfilter wird auf „Gering“ eingestellt, weil das Löschen von unnötigen Wörtern relativ einfach ist. Nach der Bestätigung der Extraktion von Einwortbenennungen wird die maximale Länge der Termini manuell von „zwei“ auf „vier„ geändert. Nach der Einstellung der maximalen Länge mit vier Wörtern wird der Qualitätsfilter auf mittel gestellt, weil es nicht so viele Vierwortbenennungen gibt (siehe Abbildungen 3.10). Die längste Mehrwortbenennung in der Datenbank ist „Amtliches Mitteilungsblatt der Hochschule Anhalt“, die aus fünf Wörtern besteht. Mehrwortbenennungen mit fünf Wörtern sind selten, und werden schon bei der Extraktion von Zweiwortbenennungen bzw. Dreiwortbenennungen in dem KWIC-Index erkannt und manuell hinzugefügt.

Bei der Termextraktion von Zweiwort- und Dreiwortbenennungen werden folgende Synonyme identifiziert:

- Zwei Wörter: weiteres Studium (Weiterstudium), Weiterführende Ausbildung (Weiterbildung)
- Drei Wörter: Anmeldung zur Prüfung (Prüfungsanmeldung), Grund der Beurlaubung (Beurlaubungsgrund).

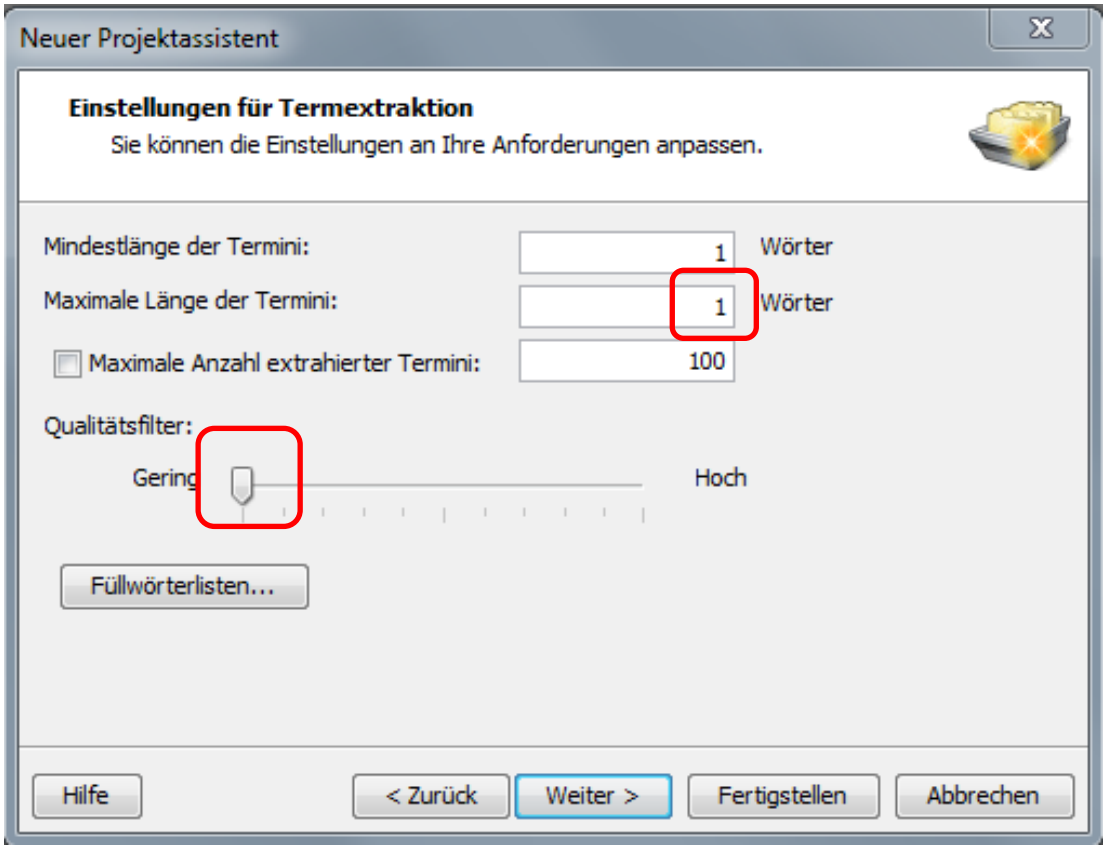


Abbildung 3.9 SDL MultiTerm 2014 Extract - Einstelllungen für die Termextraktion im Projekt

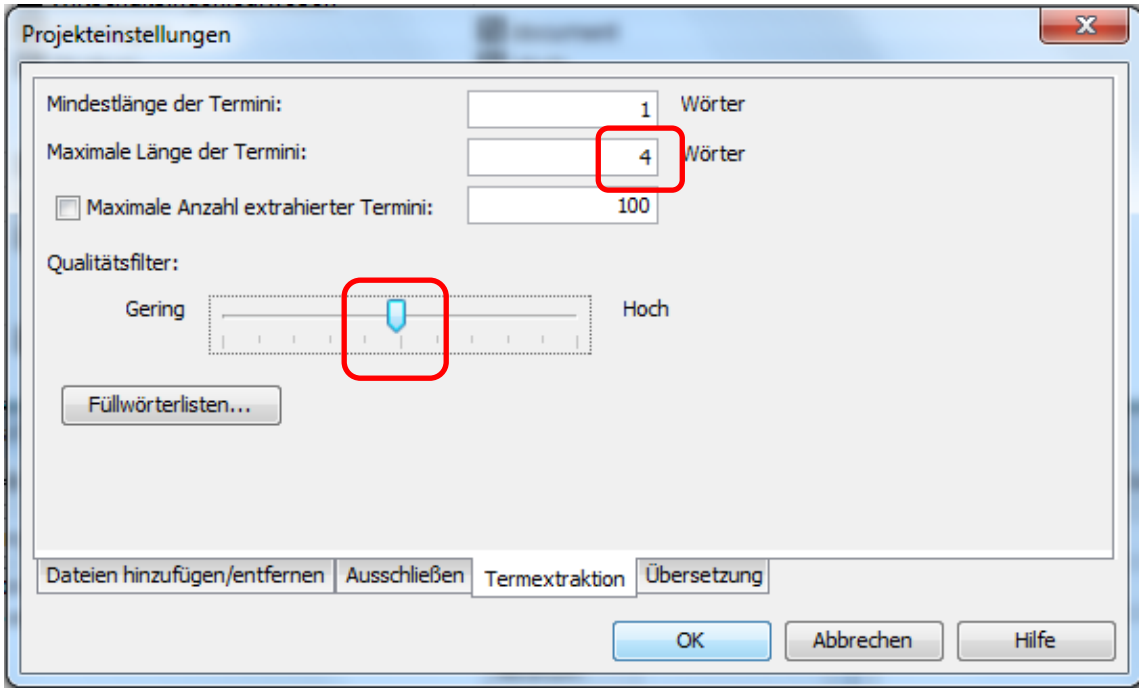


Abbildung 3.10 SDL MultiTerm 2014 Extract - Einstellungen von Terminilänge und Qualitätsfilter

Die Zahlen für maximale Übersetzungen eines Termkandidaten und die minimale Häufigkeit eines Zielkandidaten sind „fünf“ und „drei“ (siehe Abbildung 3.11). Die Qualität hat keinen großen Unterschied mit den Kombinationen der Zahl von eins bis fünf bei der zweisprachigen Extraktion. In dieser Arbeit verändert sich die Zahl hier nicht.

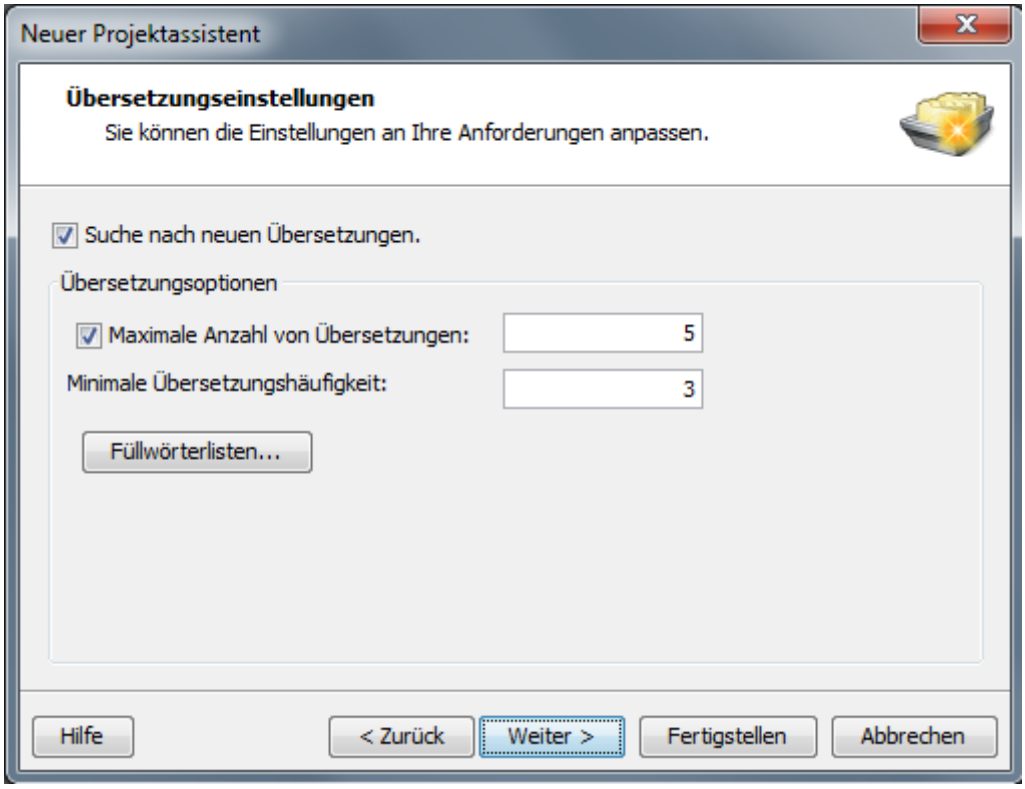


Abbildung 3.11 SDL MultiTerm 2014 Extract - Übersetzungseinstellungen

Mit SDL MultiTerm 2014 Extract ist die Extraktion sehr schnell. Es ist vorherzusehen, dass nicht alle Termini extrahiert werden. Bei der Übersetzung oder weiterer Bearbeitung kann die Termbank ergänzt und gepflegt werden. Nach dem Alignment gehen einige Fachwörter, die keine Übersetzung in den gemischten Dateien haben, verloren. Wenn die anderen Dateien diese Fachwörter enthalten, können Quellen ohne ohne Äquivalent bei der Extraktion nicht erfasst werden. Um dieses Problem zu vermeiden, sind die Ausgangstexte in die Zieltexte in MS-Word vor dem Alignment zu übertragen.

Die zweisprachige Extraktion ist problematisch, da bei der zweisprachigen Extraktion nicht nur die Informationen einer Sprache, sondern auch die Beziehungen zwischen den zwei Sprachen mit komplexen Informationen aus Begriffen und Benennungen analysiert werden sollen. Bei der automatischen Extraktion werden folgende Probleme gefunden:

- Mehrwortbenennungen, die Interpunktion oder Symbole enthalten, werden nicht erkannt. Z. B. die Bezeichnung eines Fachbereichs „Arichitektur, Facility Management und Geoinformation“ ist wegen der Interpunktion „Komma“ nicht erkennbar. Nur die Wörter „Architektur“, „Facility Management“,

„Geoinformation“ oder „Facility Management und Geoinformationen“ werden erkannt. Die ganze Bezeichnung kann nur nach dem KWIC-Index manuell hinzugefügt werden.

- SDL MultiTerm 2014 Extract hat seine eigene Stoppwortliste, deswegen werden viele allgemeine Wörter nicht extrahiert. In dieser Termbank werden sie manuell extrahiert. Ein Grund dafür ist, dass eine längere Benennung im natürlichen Sprachgebrauch beim Wiederholen häufig verkürzt wird. Die verkürzten Benennungen sind meistens allgemeine Wörter und Synonyme ihrer Vollform und treten häufiger als ihre Vollform in einem Text auf. Die Extraktion von diesen verkürzten Benennungen ist besonders wichtig für die mehrsprachige Terminologieverwaltung und die zukünftigen Übersetzungen. Es ist notwendig, die Bemerkungen für die verkürzten Benennungen hinzuzufügen, um die möglichen Probleme zu vermeiden und eine bessere mehrsprachige Terminologieverwaltung durchzuführen.

- Die Akronyme werden nicht im Feld „Akronym“ sondern im Feld „Synonyme“ eingegeben, weil die Akronyme nach dem Export nicht in SDL MultiTerm 2014 Desktop wie erwartet angezeigt werden.

- Es gibt Kodierungs-Fehler bei der TXT- bzw. DOC-Datei, so dass einige ä, ö, ü, ß durch andere Zeichen dargestellt werden. Ein möglicher Grund dafür ist, dass manche Arbeiten mit einem chinesischen System am Computer durchgeführt werden. Zeichensalat kann auch durch die Überführung von Dateiformaten in TXT-Format bzw. der Einstelllung von der Kodierung erzeugt werden.

Score	Domain	Deutsch(de)
62	<None>	Geoinformatik
86	<None>	Genehmigung der Zulassung zur Abschlussarbeit oder Kopie des Abschlusszeugnisses
81	<None>	genaue Bezeichnungen gemäß der Prüfungsordnung
75	<None>	genaue Bezeichnung
61	<None>	genannt
87	<None>	Gem溥 Landesdatenschutzgesetz stimme ich der Erfassung und Verarbeitung
41	<None>	Geltungsdauer
70	<None>	Geburtsort
41	<None>	Geburtsname
66	<None>	Geburtsland
89	<None>	Geburtsdatum
99	<None>	geben Sie den von Ihnen gewünschten Studiengang
66	<None>	geb
41	<None>	Gasthörer
84	<None>	儡 eingeschriebene Studierende und AbsolventInnen der HSA
74	<None>	Führen der Berufsbezeichnung
99	<None>	FS
63	<None>	Fremdsprache Englisch
70	<None>	freiwilliges ökologisches

Abbildung 3.12 nicht korrekt dargestellte Zeichen in MultiTerm Extract

- Der Punkt am Ende einer Abkürzung wird nicht extrahiert, z. B. der Terminus „Matrikel-Nr.“.

- Ein Teil der Wortgruppen (Mehrwortbenennungen mit Ellipse) kann nur bei der Postedition extrahiert werden. Ein Beispiel dafür ist, dass das Wort „Grundwehr“ anstatt des Wortes „Grundwehrdienst“ aus „Grundwehr- und Zivildienst“ extrahiert wird.

- Die Schreibvariante Großbuchstaben, z. B. HOCHSCHULE ANHALT (FH) wird nicht erkannt.

- Die englischen Wörter, die in deutschen Dateien existieren, werden nicht extrahiert.

- Wie in Abbildung 3.13 angezeigt ist der Kontext eines Termkandidaten manchmal nicht vollständig. Die Modulbezeichnung „Internat. Business (IBS)-engl. Zweig“ wird als „Business (IBS)“ aus dem Kontext „Business (IBS)-engl.“ extrahiert. Links der Abbildung ist der Kontext in der PDF-Datei.

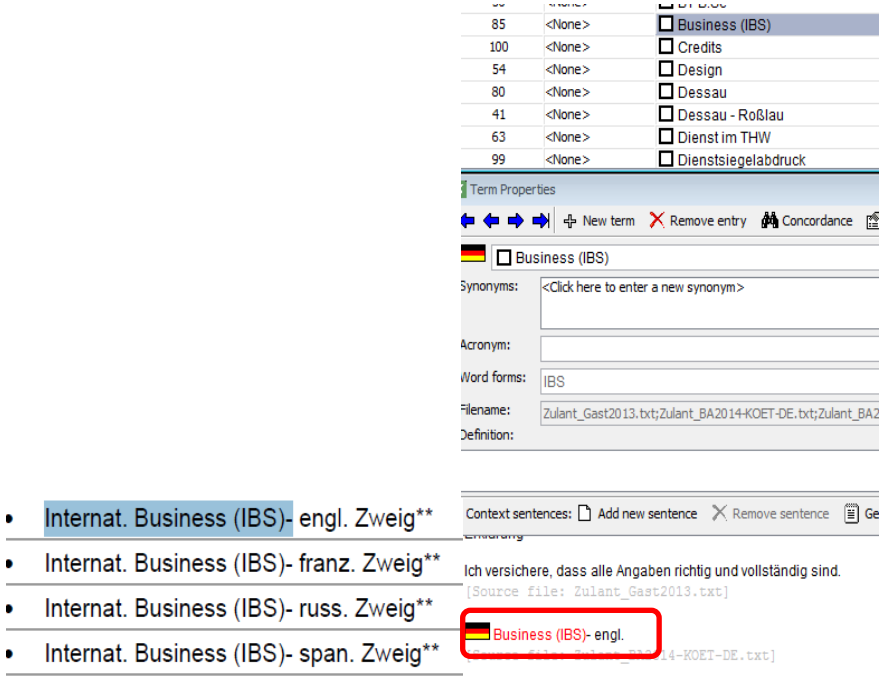


Abbildung 3.13 PDF-Datei (links) und SDL MultiTerm 2014 Extract (rechts)-Vollständigkeit des Kontexts

- Das Werkzeug macht durch die unscharfe Such auch Fehler. Es ist nicht typisch, aber es kommt vor. Wie in der folgenden Abbildung 3.14 angezeigt, wird das Wort „Student“ extrahiert. Aber in dem Kontext wird nicht nur das Wort „Student“ und sein Kontext sondern auch das Wort „Stunden“ mit seinem Kontext mit Rot gezeichnet.

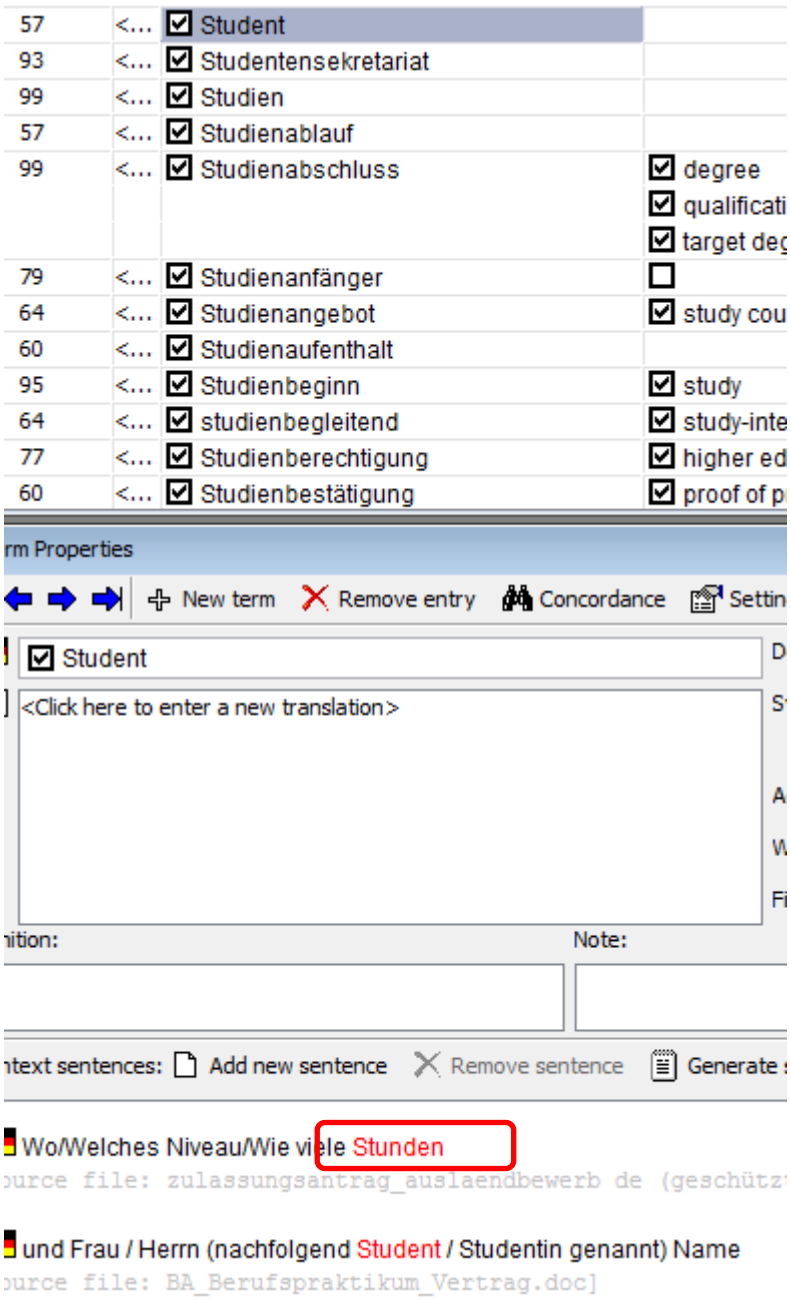


Abbildung 3.14 Zuordnungsfehler aus SDL MultiTerm 2014 Extract

- Die Sortierung von Wörtern, die mehrere Bedeutungen haben, ist schwer. Die anderen Bedeutungen eines Wortes können nur in dasselbe Feld eingetragen werden, sonst wird eine Warnung bei Duplikaten angezeigt, d. h. die folgenden zwei Begriffe können nur in einem Feld eingetragen werden. (Abbildung 3.15)

Abbildung 3.15 Behandlung von Polysemie

- Wenn die Ergebnisse als TXT-Datei exportiert werden, können die Synonyme nicht erkannt werden. Beim Export in eine MultiTerm-Datenbank werden die Synonyme entsprechend extrahiert und angezeigt. Ein Nachteil davon ist, dass alle Synonyme vor dem Export entsprechend zugeordnet werden müssen. Das ist aber schwer auf einmal zu erledigen.

Es gibt auch Vorteile im Vergleich zur manuellen Extraktion.

- Die meisten Singularformen werden erkannt, typischerweise beim Wort mit der Endung -e oder –en. Z.B das Wort „Bearbeitungsvermerk“ mit der Endung -e wird aus dem Kontext „Bearbeitungsvermerk**e** des Prüfungsamtes“ extrahiert.

- Fast alle Kontextbeispiele eines Terminus werden angezeigt.

- Die Häufigkeit der Erscheinung eines Terminus wird im Feld nis“ („Score“) angezeigt.

- Die verschiedenen Schreibvarianten können gut erkannt werden, die bei der manuellen Extraktion missachtet werden. Ein gutes Beispiel wird in Abbildung 3.16 angezeigt.

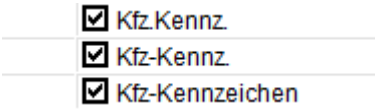


Abbildung 3.16 Erkennung von Schreibvarianten

3.2.3 Vergleich der zwei Extraktionsmethoden

Die Extraktion in MS-Excel ist zeitaufwändiger und kann bei großen Mengen von Ausgangsmaterialien durch Übermüdung oder wegen der fehlenden Kenntnisse in einem bestimmten Fachgebiet Fehler erzeugen. Mit SDL MultiTerm 2014 Extract wird Zeit eingespart, aber es entstehen gleichzeitig Nachteile bei

der Vollständigkeit und Richtigkeit. Für große Mengen von Ausgangsmaterialien wird SDL MultiTerm 2014 Extract zur Extraktion bevorzugt.

In Excel können die Termkandidaten besser als die in SDL MultiTerm 2014 Extract angeordnet werden. Für das Hinzufügen und die Bearbeitung von Zusatzinformationen hat MS-Excel auch eigene Vorteile. Die Bearbeitung von Termkandidaten sowie ihre Zusatzinformationen ist in MS-Excel flexibler, während die Termkandidaten in SDL MultiTerm 2014 Extract nach dem Import in SDL MultiTerm 2014 Desktop mehr Bearbeitung benötigen.

In dieser Arbeit werden die Termkandidaten in Excel extrahiert. Die Extraktion mit SDL MultiTerm 2014 Extract dient zum Vergleich mit der Extraktion in MS-Excel und als Ergänzung und Überprüfung.

3.3 Erstellen und Erweiteren einer Terminologiedatenbank

In Terminologiedatenbanken werden die Termkandidaten überprüft, die Termini und die mehrsprachigen Äquivalente weiter erfasst und vereinheitlicht. Die Benennungen und die äquivalenten Termini werden nach der Überprpüfung festgestellt und freigegeben. Diese Bearbeitungen können entweder in MS-Excel oder in SDL MultiTerm 2014 Dektop durchgeführt werden. In dieser Arbeit werden die oben genannten Bearbeitungen meistens in MS-Excel durchgeführt, um eine flexible Arbeit durchzuführen. Die Termbankdefinition wird von der Projektkoordinatorin, Frau Prof. Dr. Uta Seewald-Heeg, vorher zur Verfügung gestellt. Allgemeine Informationen und beschreibende Felder sind wie folgt strukturiert.

Languages
Chinese
English
German

Entry Structure		
	Mandatory	Multiple
Eintragsebene		
Sachgebiet		•
Definition		•
Definitionsquelle		•
Zugriffsdatum		•
Kommentar		•
Sprachebene		
Termebene		
Quelle		•
Zugriffsdatum		•
Wortart		•
Genus		•
Termtyp		•
Status		•
Kontextbeispiele		•
Quelle Kontext		•
Kommentar		•

Descriptive Fields			
Name	History	Type	Picklist Values
Definition		Text	
Definitionsquelle		Text	
Genus		Picklist	maskulin | feminin | neutral
Kommentar		Text	
Kontextbeispiele		Text	
Quelle		Text	
Quelle Kontext		Text	
Sachgebiet		Picklist	Allgemein | Wissenschaft und Hochschule
Status		Picklist	bevorzugt | zugelassen | verboten | abgelöst | wird überprüft
Termtyp		Picklist	Abkürzung | Akronym | Vollform | Kurzform | Phraseologische Einheit
Wortart		Picklist	Substantiv | Adjektiv | Verb | Adverb | Eigenname | Sonstiges
Zugriffsdatum		Text	

Abbildung 3.17 SDL Trados Studio 2014 - Termbankdefinition

3.3.1 Konvertieren terminologischer Daten

Es gibt zahlreiche Programme zur Terminologieextraktion und zur Erstellung eines Terminologiebestands. Um den Datenaustausch zwischen den verschiedenen Systemen reibungslos durchzuführen, werden die Daten in ein einheitliches Format konvertiert. Auch die Schnittstellen werden für den Datenaustausch eingerichtet. Dafür werden einige Austauschformate, wie zum Beispiel

MARTIF (Maschine-readable Terminology interchange Format) und TBX (TermBase eXchange) entwicklet.

Durch den Einsatz von SDL MultiTerm 2014 Convert werden die Struktur der Excel-Datei in XDT-Datei und der Inhalt der Termbank in XML-Datei angelegt. Die in MS-Excel eingetragenen Werte (Picklist Values) werden automatisch angezeigt. Die fehlenden Werte müssen nach den in Abbildung 3.17 angezeigten Werten beim Import in die Zieldatenbank hinzugefügt werden.

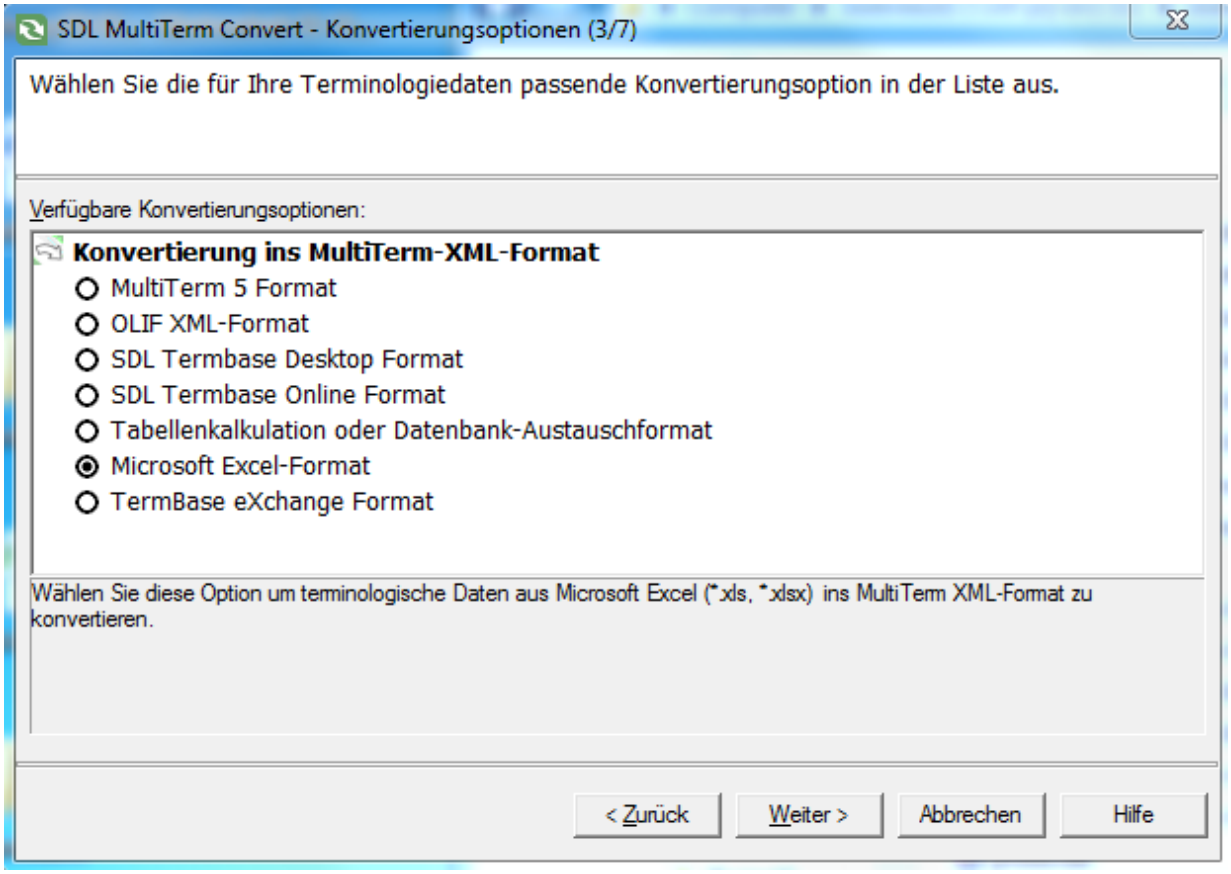


Abbildung 3.18 Konvertierungsoptionen aus SDL MultiTerm Convert

3.3.2 Importmöglichkeiten

Um die Datenbank in MultiTerm zur Verfügung stellen zu können, müssen die verwendeten Methoden eine Importmöglichkeit besitzen.

Für die einfache Methode mit Excel:

Eine Voraussetzung für den Import ist, dass die Struktur der Excel-Datei mit den Attributen der Termini identisch übereinstimmt. Das heißt, dass die Kopfzeile der Wortliste in der Excel-Datei und die Struktur der Datenbank gleich sind. In SDL MultiTerm 2014 Desktop werden die Struktur oder Definition der Terminologiedatenbank in der XDT-Datei und der Inhalt der Terminologie in eine XML-Datei importiert. Um die Synonyme richtig zuzuordnen, müssen die Einträge über Entragsnummern synchronisiert und zusammengeführt werden. Die Abbildungen 3.19 - 3.21 zeigen die Anweisungen der Auswahl der Optionen beim Import an.

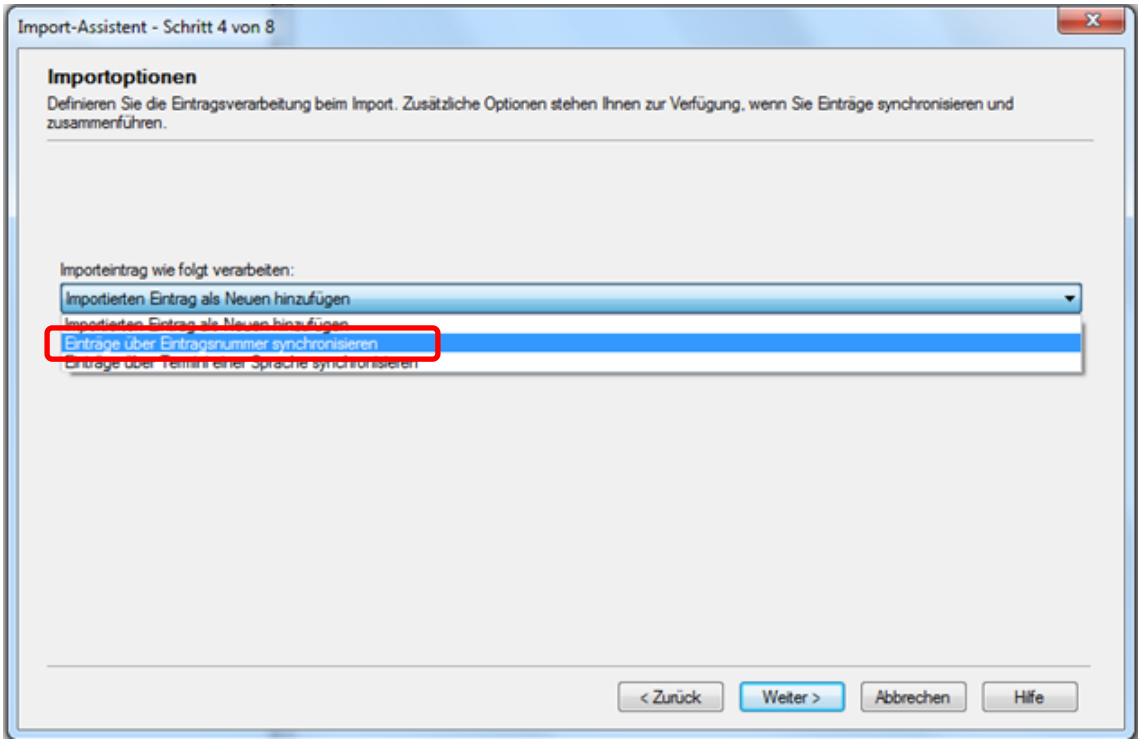


Abbildung 3.19 SDL MultiTerm 2014 Desktop - Einstellung von Importeintrag

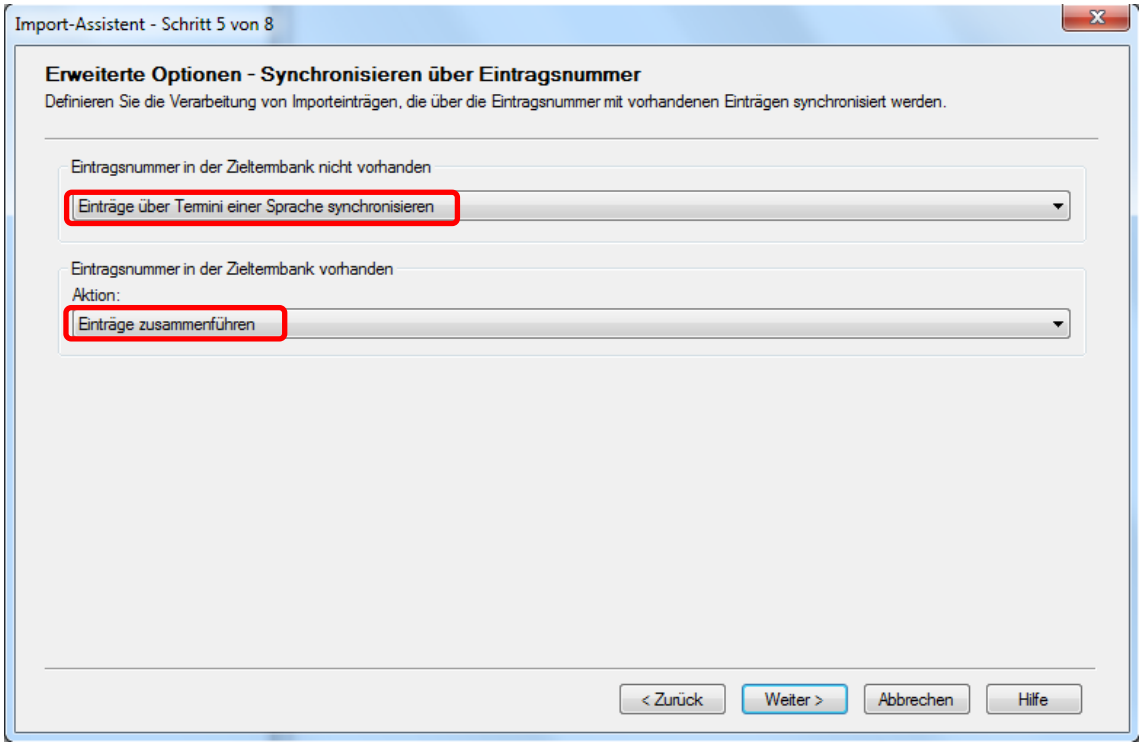


Abbildung 3.20 SDL MultiTerm 2014 Desktop - Synchronisieren über Eintragsnummer

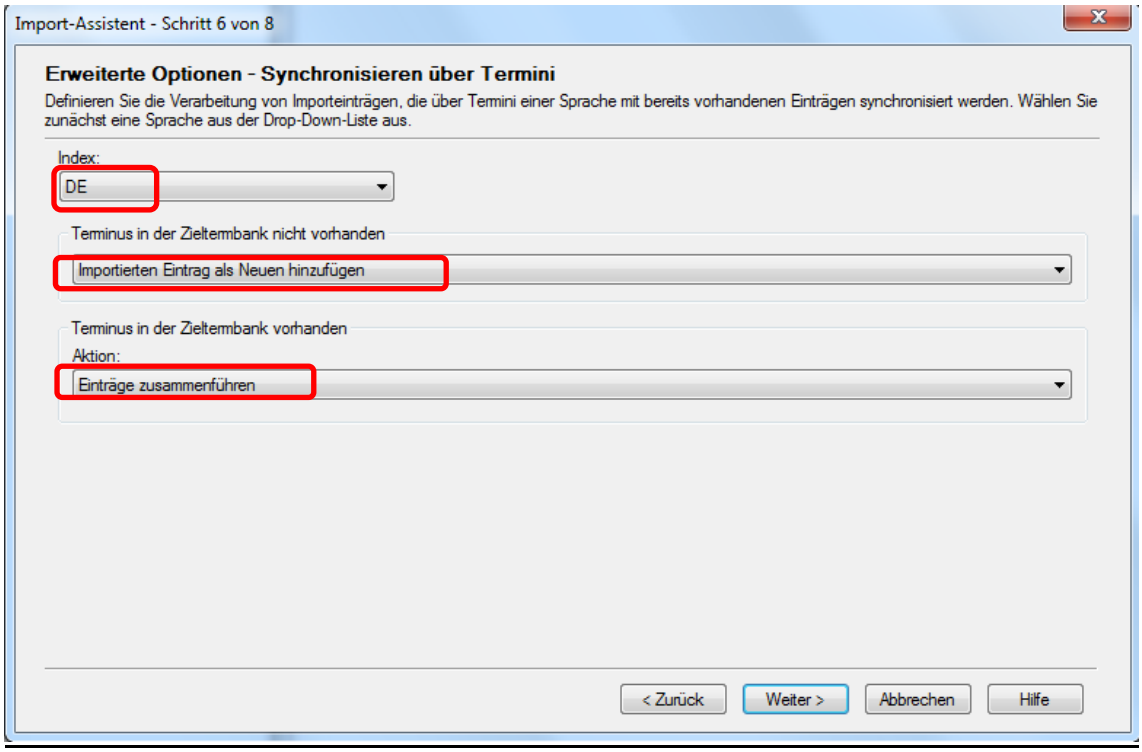


Abbildung 3.21 SDL MultiTerm 2014 Desktop - Synchronisieren über Termini

Für den Einsatz des Programms SDL MultiTerm 2014 Extract:

Um eine bessere Qualität der Termini zu bekommen, kann die Terminologieliste zuerst in eine TXT-Datei exportiert und dann in einer MS-Excel bearbeitet werden. Mithilfe der oben genannten Methode mit MultiTerm Convert wird das Dateiformat konvertiert und in MultiTerm importiert.

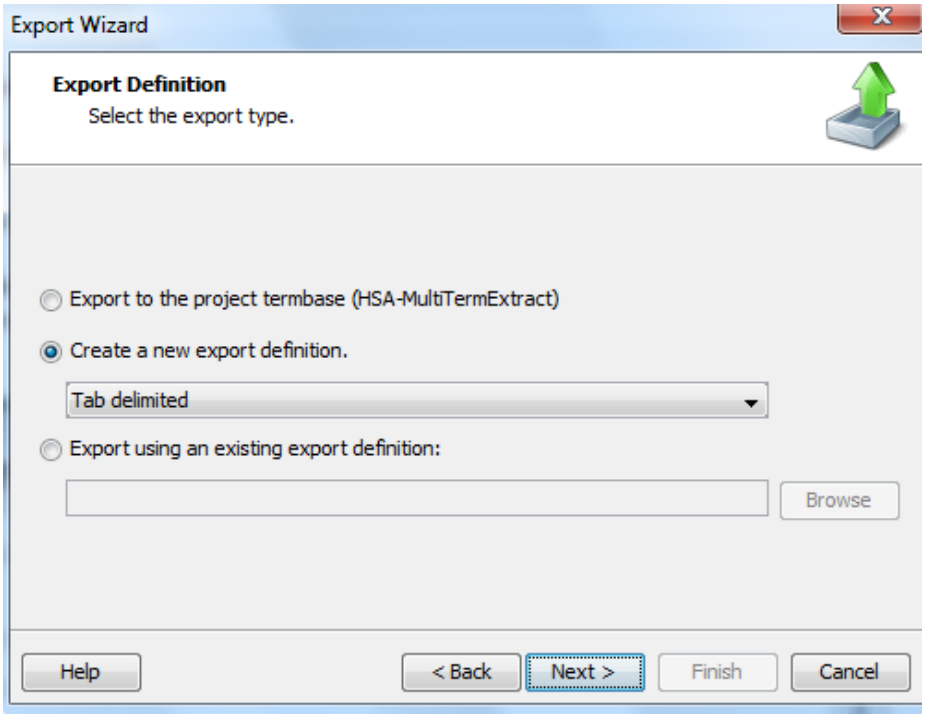


Abbildung 3.22 Export Definition

3.3.3 Exportmöglichkeiten

SDL MultiTerm 2014 Desktop bietet viele Exportmöglichkeiten an, zum Beispiel eine zweisprachige Wortliste oder ein Wörterbuch mit Definitionen.

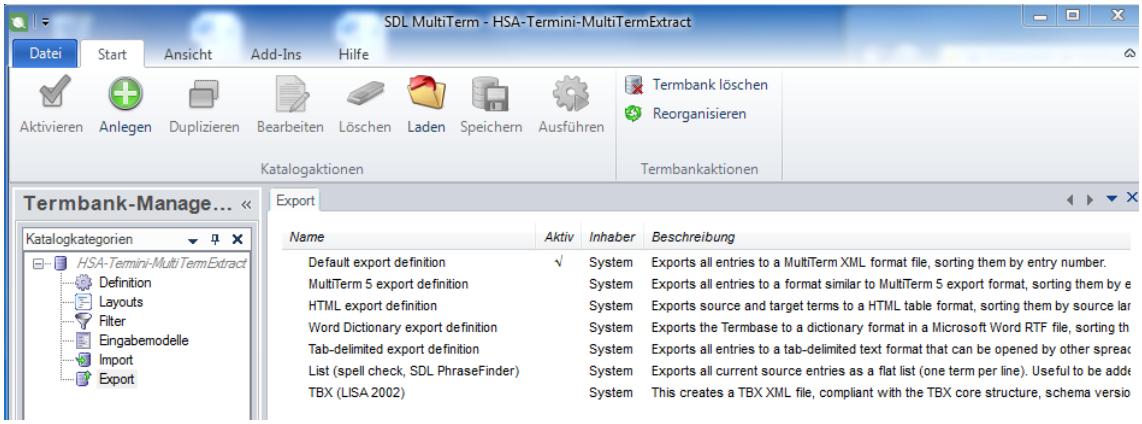


Abbildung 3.23 SDL MultiTerm 2014 Desktop - Exportmöglichkeiten

3.4 Qualitätssicherung

Ähnlich wie bei Übersetzungen sollten Terminologiedatenbankinhalte auch nach dem 4-Augen- oder 6-Augen-Prinzip überprüft werden. Grundsätzlich sollen folgende Merkmale überprüft werden:

- Richtigkeit (fachliche Korrektheit von Benennungen, Richtigkeit von Definitionen, Mehrwortbenennung, inhaltliche sowie Formatfehler)
- Vollständigkeit (fehlende Inhalte)
- Synonyme oder Duplikate
- Darstellung von Quellen sowie der entsprechenden Zugriffsdaten
- Funktionalität der Hyperlinks und der Querverseise

3.5 Abstimmung und Freigabe

Bei der Abstimmung und Freigabe eines Prozesses wird durch den Status entschieden, welche Benennungen eines Begriffs als „bevorzugt“, „zugelassen“ oder „verboten“ verwendet werden sollen. Diese Entscheidung basiert auf den Kriterien zur Term-Standardisierung (siehe Kapitel 2.2) oder des Terminologieleitfaden. Die Unklarheiten eines Begriffs müssen mit Fachleuten diskutiert werden.

3.6 Aufbereitung und Bereitstellung

Die Zielgruppen dieses Projektes sind die Mitarbeiter von ASA und die anderen hochschulintern Mitarbeiter und die Studierenden, denen kein MultiTermDesktop zur Verfügung steht. Um die Termini einwandfrei verwenden zu können, ist die Einstellung einer webbasierten Online-Version erforderlich. Dabei ist nur ein einfacher Zugang zu den Daten notwendig. Dadurch können die Benutzer dies zur Terminologiearbeit zum Nachschlagen und Überprüfen benutzen.

4 Alternative Methode zur Extraktion von Termini

In dieser Arbeit wird mit dem Translation-Memory-Werkzeug SDL Trados Studio 2014 als eine alternative Methode zur Extraktion von Termini demonstriert. Nach der Analyse von Dokumenten mit Trados Studio gibt es viele sich wiederholende Wörter bzw. Segmente in den Ausgangsmaterialien. Die Wörter der gemischten Dateien sind mehr als ein Viertel der deutschen Dateien. Die sich wiederholenden Wörter bzw. Segmente sind etwa zwei Drittel der gesamten Datei (siehe Kapitel 3.1). Zur Verdeutlichung werden nur Termini aus den sechs gemischten Dateien extrahiert.

Durch Markierung von Termini in der Ausgangssprache und in der Zielsprache, bei der bilinugalen Extraktion oder durch Markeirung von Termini in der Ausgangssprache oder bei der einsprachigen Terminologieextraktion werden die Termini in die selektierte Terminologiedatenbank hinzugefügt. In dieser Weise können neben den Benennungen auch die zusätzlichen Informationen eingegeben werden. Die Voraussetzung dafür ist, dass eine leere Termbank bei der Erstellung eines neuen Projektes mit einer vordefinierten Eintragsstruktur hinzugefügt wird. Die Bearbeitungen der Termbank in MultiTerm 2014 Desktop erscheint in einem eigenen Fenster in Trados Studio.

Besonders ist die Extraktion von zweisprachigen Dateien. Der Ausgangstext und der Zieltext sind parallel, was sehr gut für die zweisprachige Extraktion ist. Bei der manuellen Extraktion oder mithilfe von anderen Extraktionswerkzeugen wird der Gesichtpunkt nach der Gewohnheit der Terminologen auf der Ausgangssprache (hier Deutsch) eingerichtet: Aber in Trados Studio kann der Zieltext auch als Gesichtpunkt angewendet werden. Ein Beispiel wird in Abbildung 4.1 angezeigt, wo die Termkandidaten bei der manuellen Extraktion schwer zu identifizieren sind.

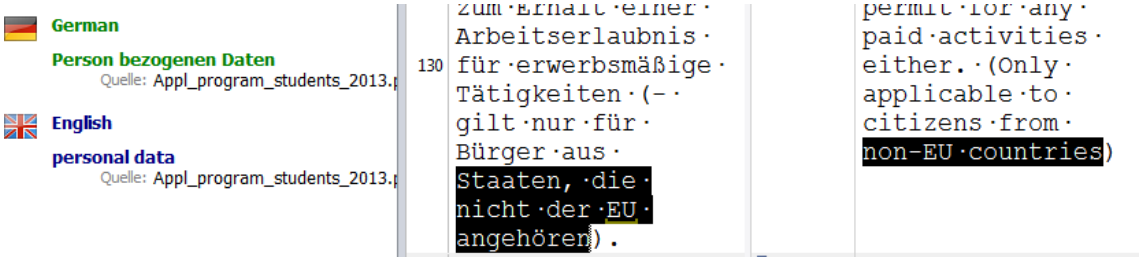


Abbildung 4.1 Beispiel der Extraktion in Trados Studio

Das Ergebnis der Extraktion kann nur in MultiTerm stehen. Es ist anderes als in SDL MultiTerm 2014 Extract, das die Termkandidaten noch in Textformat exportiert. Mit Trados Studio müssen zuerst die Termkandidaten und die geeigneten Quellen in die Termbank hinzugefügt werden. Dann werden die anderen Informationen in die Termbank ergänzt werden. Es gibt auch eine andere Möglichkeit, die Kontextbeispiele und die Quellen der Kontextbeispiele bei der Extraktion direkt in die Termbank hinzuzufügen. Gleichzeitig müssen alle anderen Informationen ergänzt werden. Ansonsten wird die Reihenfolge der Eintragsfelder wie Wortart, Status, Termtyp usw. durcheinander dargestellt. (siehe Abbildung 4.2)

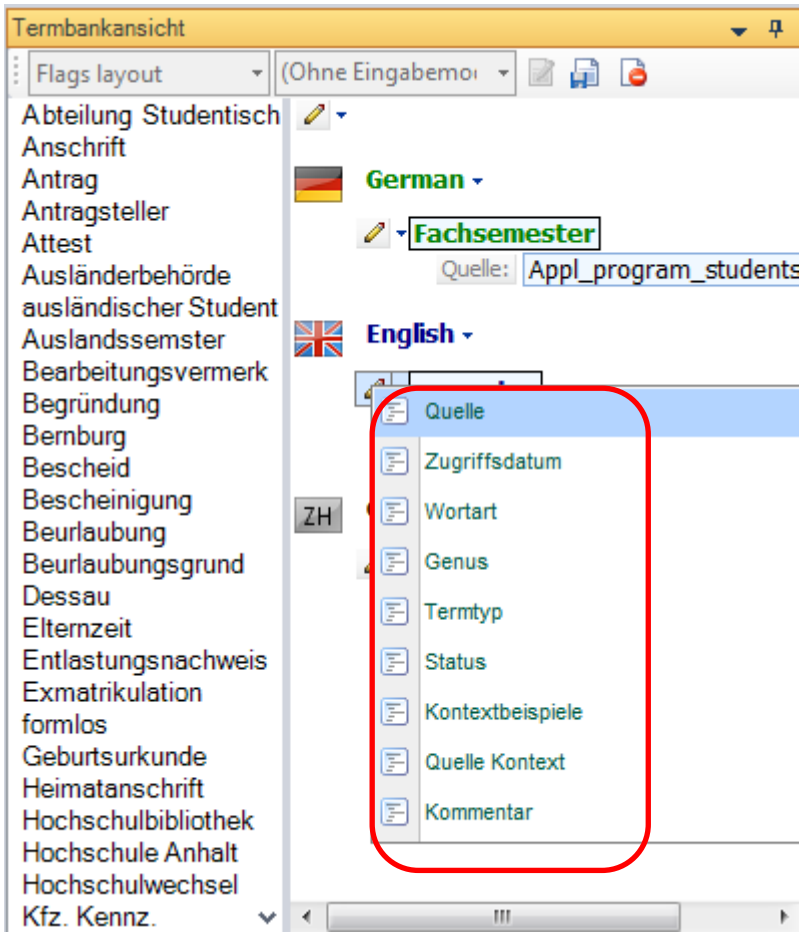


Abbildung 4.2 SDL Trados Studio 2014 – Eintragsfelder in Termbankansicht

Beim Hinzufügen eines Terminus geht es bei der Markierung einer deutschen Benennung (Ausgangssprache) oder einer deutschen und gleichzeitig einer englischen Bezeichnung (egal in welcher Zeile), aber nicht nur einer englischen Bezeichnungen (Zielsprache):

Als eine alternative Methode zur Terminologieextraktion hat Trados Studio folgende Vorteile und Nachteile:

- Viele Synonyme können gut sortiert werden, typischerweise bei einem deutschen Terminus mit mehr englischen Äquivalenten oder bei identischer englischer Übersetzung von deutschen Synonymen. Die Abbildung 4.3 zeigt ein Beispiel dazu an. Die beiden deutschen Benennungen „erstmalige Immatrikulation“ und „Erstimmatrikulation“ werden gleich mit „first admission“ übersetzt. Es wird beim Hinzufügen gefragt, ob der Eintrag in Bearbeitung genommen werden soll (siehe Abbildung 4.4).

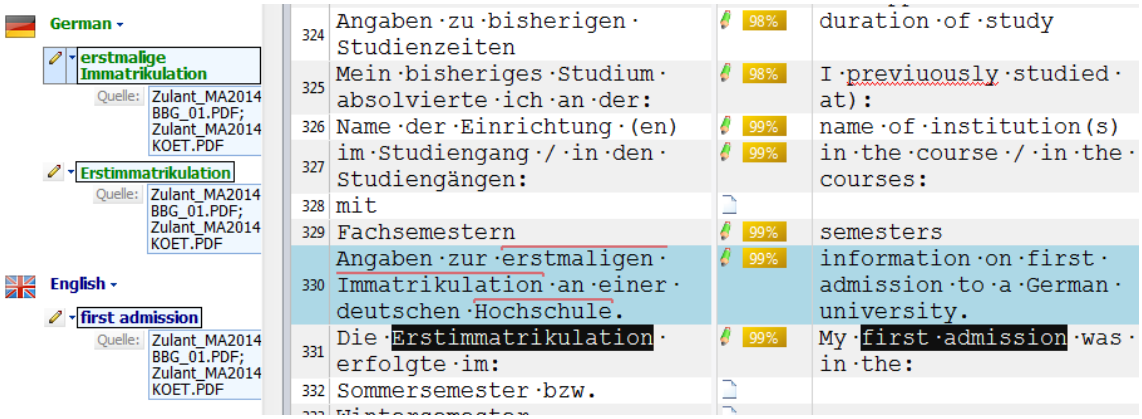


Abbildung 4.3 SDL Trados Studio 2014 - Eintrag von Synonymen in Termbankansicht

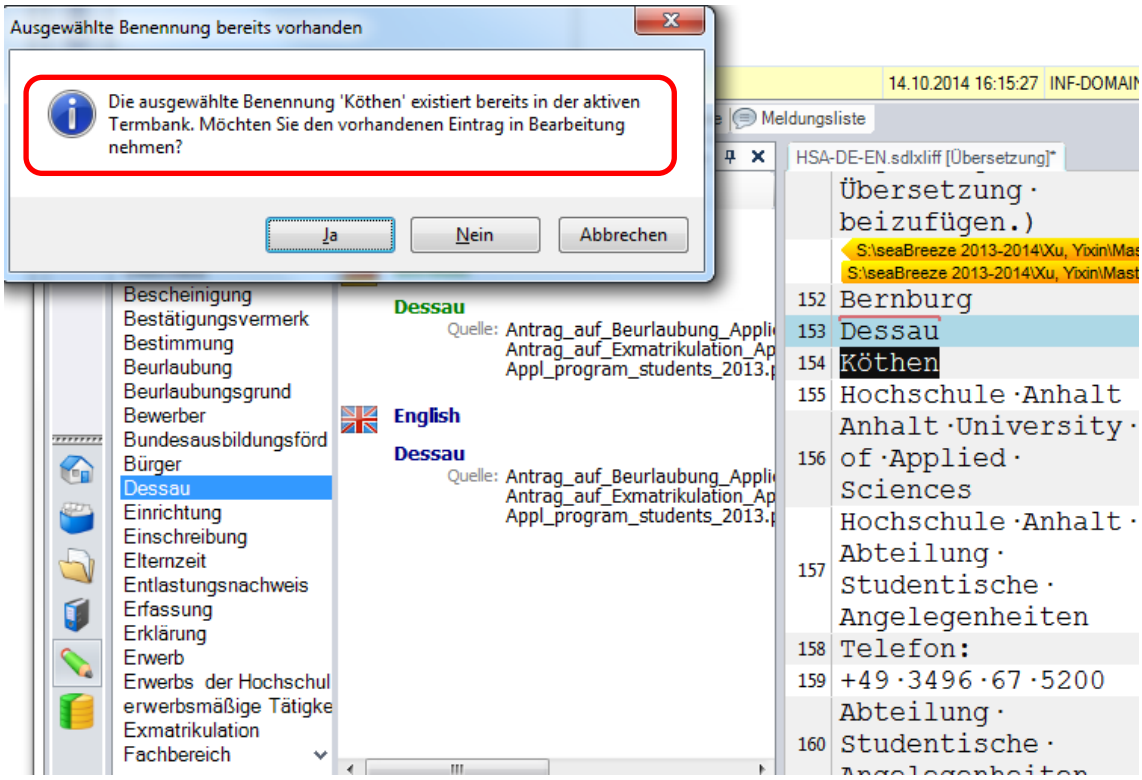


Abbildung 4.4 SDL Trados Studio 2014 - Befragen bei der Wiederholungen in Termbankansicht

- Kontextbeispiele sind notwendig. Beim mehrmaligen Eintragen der Quelle eines Begriffs kann nach den Kontextbeispielen entschieden werden, ob der neue Eintrag als Synonym (meistens Quasisynonyme) eines Termkandidaten eingetragen werden soll.

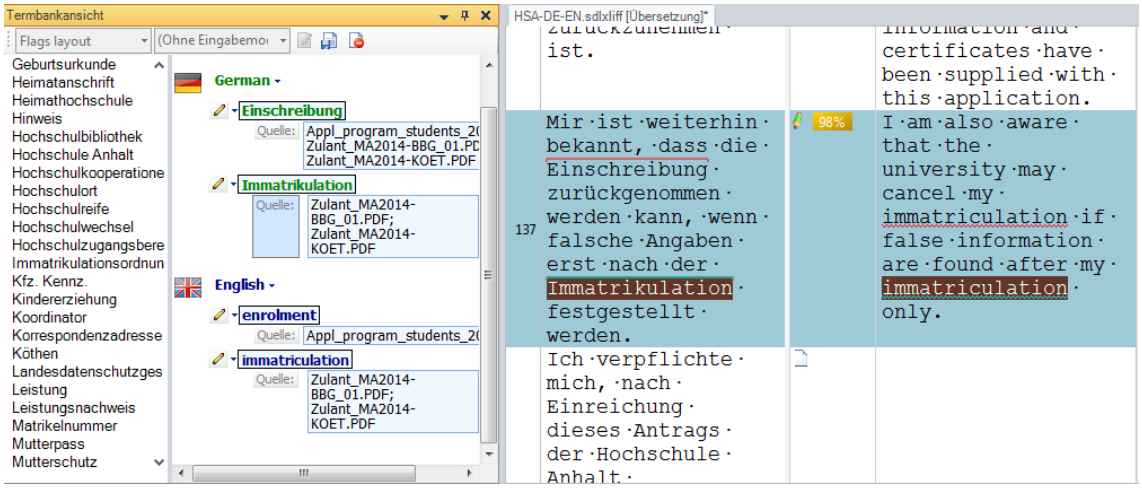


Abbildung 4.5 Quasisynonyme

- In Abbildung 4.6 wird die Quelle der Ausgangssprache durch die roten Rechtecke 1 markiert. Im roten Rechteck 2 gibt es die Anweisungen für die Quellen der Zielsprache. Es ist unmöglich, hier alle Quellen einmal darzustellen, da die Quelle hier die der Segmente ist. Einige Quellen der Ausgangssprache sind nicht vollständig. Im Fenster nung“ zeigt das rote Rechteck 3 die eingetragten Termini an.

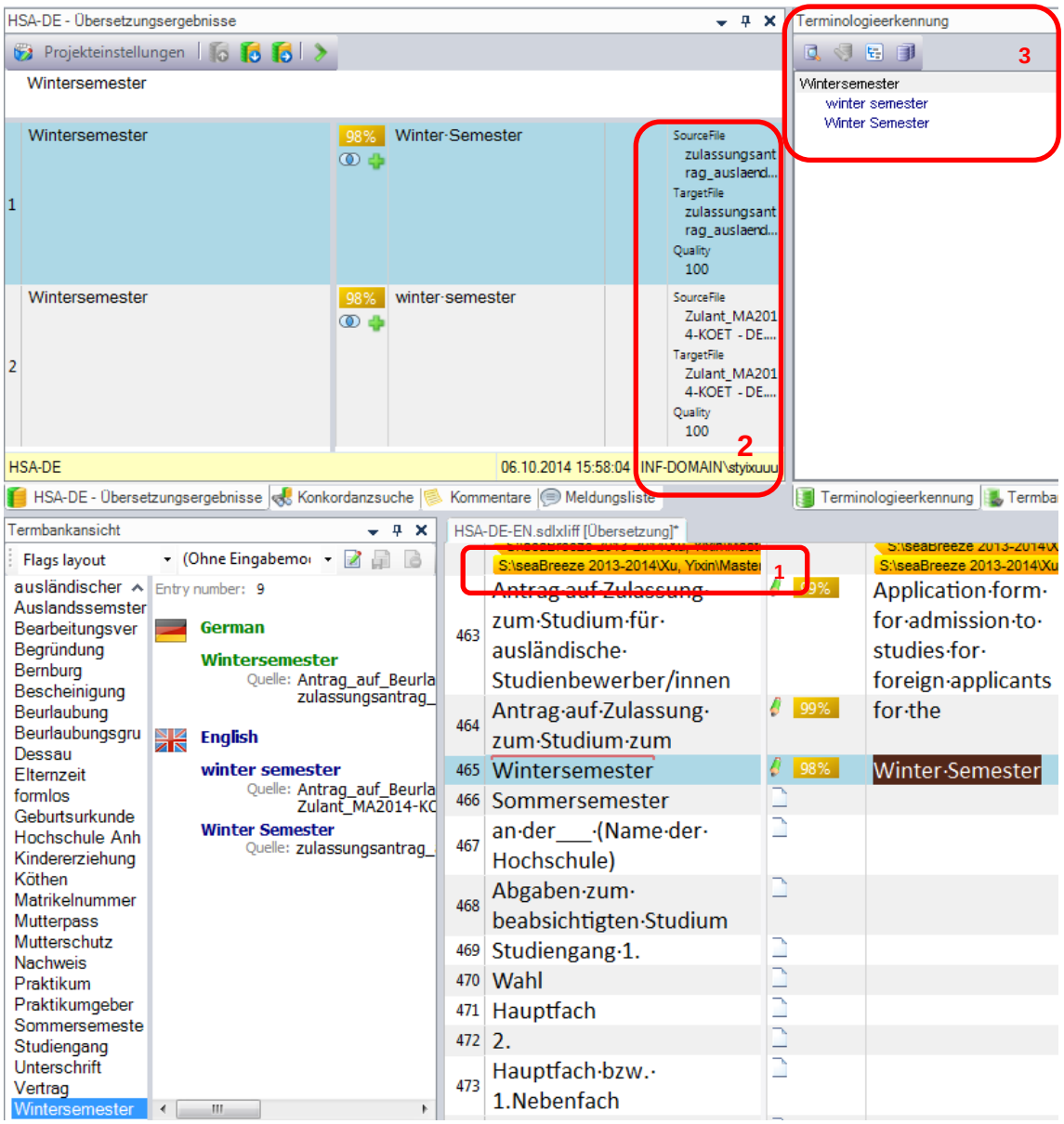


Abbildung 4.6 Quelleanzeige in Trados Studio

- Die Segmente werden durch Satzzeichen am Ende eines Satzes getrennt. Enthält eine Abkürzung zwei Punkte und steht gleichzeitig nach dem ersten Punkt ein Leerzeichen, wird diese Abkürzung in zwei Segmente getrennt (siehe Abbildung 4.7). Solche Abkürzungen sind in Trados Studio schwer zu bemerken.

Abbildung 4.7 Extraktionsprobleme in Trados Studio

5 Evaluation

In dieser Arbeit wird MS-Excel bei der Extraktion verwendet. Die Werkzeuge SDL MultiTerm 2014 Extract und SDL Trados Studio 2014 werden auch genutzt, um die Qualität der Extraktion zu gewährleisten. Das Ergebnis der Termextraktion von SDL MultiTerm 2014 Extract ist eine gute Referenz zur Überprüfung der Termini, z. B. der Vergleich zwischen den unnötigen Benennungen und Mehrwortbenennungen. In SDL MultiTerm 2014 Extract gibt es eine „Füllwörterliste“, mit der einige unnötige Benennungen ausgeschlossen werden können. Das „Ergebnis“ in SDL MultiTerm 2014 Extract bedeutet die Häufigkeit des Termkandidaten. Mit der Funktion „als neuer Terminus hinzufügen“ in Trados Studio können auch viele Synonyme, die die gleiche Übersetzung haben, erkannt und viele Quellen hinzugefügt werden. So ist es möglich, die Synonyme zu überprüfen und die Vollständigkeit der Quellen von Termini zu verbessern.

Mit maschineller Extraktion wird die Termkandidatenliste sehr schnell erzeugt. Es dauert nur ein paar Sekunden. Die gesamte Bearbeitungszeit ist abhängig von der Textlänge. Die Bearbeitungszeit ist nach der Textlänge nur halb solang oder sogar kürzer als bei der manuellen Extraktion.

Nach Eckstein 2009[33] werden die Evaluierungen oder die sogenannten Vergleichskriterien für Terminologieextraktionsprogramme wie folgt beschrieben:

- Qualität der Extraktion von Termkandidaten
- Möglichkeiten des Datenaustauschs
- Behandlung von Benennungen
- Behandlung von Synonymie
- Behandlung von Zusatzinformationen
- Unterstützung von Sprachen und Mehrsprachigkeit.

[33] [Eckstein 2009: Seite 110-113]

5.1 Qualität der Extraktion von Termkandidaten

Hier geht es um die Kriterien „Silence“ und „Noise“ (siehe Kapitel 2.4.1). Es ist schwer zu definieren, ob ein Terminus wichtig ist. Die Stoppwortliste spielt dabei eine große Rolle. Aufgrund des Einsatzes des „Qualitätsfilters“ und von „Füllwörterlisten“ ist die Qualität der Termkandidaten in einsprachiger Termextraktion aus SDL MultiTerm 2014 Extract nur ein bisschen schwächer als bei den anderen zwei Methoden. Nur einige relevante Termini werden nicht extrahiert.

5.2 Möglichkeiten des Datenaustauschs

Die Unterstützung von verschiedenen Dateiformaten spielt eine große Rolle beim Import. Im Vergleich zu den anderen Werkzeugen ist SDL MultiTerm 2014 Extract hier relativ schwach, weil es keine PDF-Dateien, sondern nur ältere Word-Versionen unterstützt. Ein weiterer wichtiger Punkt ist, in welchem Format die extrahierten Daten exportiert werden und ob die entstehende Datei in eine vorhandene Terminologiedatenbank importiert werden kann. Die drei verwendeten Methoden haben kein Problem, in SDL MultiTerm 2014 Desktop importiert zu werden. MS-Excel ist flexibeler beim Datenaustausch als die anderen zwei Methoden.

5.3 Behandlung von Benennungen

Es geht darum, ob ein Terminologieextraktionsprogram die Grundform eines Termkandidaten erkennt oder die angezeigten Termkandidaten auf ihre Grundform zurückführen kann. Dazu wird die Nachbearbeitung vereinfacht und Zeit eingespart. SDL Trados Studio 2014 erkennt den Numerus. Aber beim Zurückführen von Plural in die Grundform ist der Plural mit allen drei Methoden manuell nachzubearbeiten. SDL MultiTerm 2014 Extract hat den Vorteil, dass bei Duplikaten oder Wiederholungen eine Warnung erscheint.

5.4 Behandlung von Synonymie

Die Behandlung der Synonymie in Excel ist flexibel. Auch SDL Trados Studio können Synonyme gut sortiert werden, typischerweise bei einem deutschen Terminus mit vielen englischen Äquivalenten oder bei identischer englischen Übersetzung von deutschen Synonymen. In SDL MultiTerm 2014 Extract können Synonyme hinzugefügt werden. Wegen der Warnung von Wiederholungen können die anderen Bedeutungen einer Benennung nicht mehr bearbeitet werden.

5.5 Behandlung von Zusatzinformationen

Mit Excel können Zusatzinformationen beliebig hinzugefügt werden. Das Hinzufügen von Zusatzinformationen ist in SDL Trados Studio 2014 möglich, während es in SDL MultiTerm 2014 Extract mangelhaft ist. Obwohl viele Zusatzinformationen in SDL MultiTerm 2014 Extract eingetragen werden können, werden nur Synonyme in SDL MultiTerm 2014 Desktop angezeigt.

5.6 Unterstützung von Sprachen und Mehrsprachigkeit

Grundsätzlich unterstützt ein statistisches Programm fast alle Sprachen. Es ist wichtig für die Evaluierungen eines Terminologieextraktionsprogramms, ob es nur einsprachige oder auch zweisprachige Terminologieextraktion unterstützt.

6 Resümee

Der Unterschied zwischen allgemeinsprachlichen und fachsprachlichen Termkandidaten ist schwer zu definieren. Die Fachlichkeitsgrade hängen von der Erfahrung der Terminologen oder der Fachleute ab. Ein entscheidendes Merkmal zur Differenzierung von der Gemeinsprache ist, dass die Fachsprache auf ein Fachgebiet gerichtet ist. z. B. Semester, Absolvent usw. im Bereich Wissenschaft und Hochschule. Die Referenzen, die zur Auswahl der Extraktion von einigen Wörtern in dieser Arbeit genutzt werden, sind das DAAD-Wörterbuch und das Deutsch-Chinesische Universitätswörterbuch. Die Erweiterung über die Behandlung für die Quasisyononyme (Oberbegriff und Unterbegriff) ist eine schwere aber notwendige Arbeit.

Probleme bei der Extraktion:

Einsprachige Extraktion

a. Behandlung von Sonderzeichen wie Klammern, Bindestriche und Schrägstrichen: das am häufigsten verwendete Sonderzeichen ist der **Schrägstrich**, der die gleiche Bedeutung der zwei Wörter oder eine alternative Auswahl bedeutet. Dabei ist es bei manchen nicht einfach zu entscheiden, ob sie als Synonyme eingetragt werden sollen.

b. Zu viele **Quasisynonyme**: Es gibt viele Quasissynonyme, die durch Auslassen eines Wortbestandteils (Reduktionsvarianten) erzeugt werden, z. B. „Semester“ und „Hochschulsemester“.

c. Zu viele Abkürzungen bzw. verkürzte Formen (**Ellipse**): In Formularen werden Benennungen häufig anders als in Fließtexten geschrieben und verwendet. Ellipsen treten wegen des begrenzten Platzes oder zum Formulieren von Anweisungen oder Zielangaben in Formularen auch häufig auf. In dem Ausgangstext gibt es beispielsweise eine Ellipse „Art der Arbeit: Bachelor, Master, Diplom“, deren Vollform „Art der Abschlussarbeit: Bachelorarbeit, Masterarbeit, Diplomarbeit“ ist. Eine Frage

dazu ist, ob die Benennungen „Bachelor“ und „Bachelorarbeit“ als Quasi-synonyme eingetragen werden sollten. Eine Benennung in Formularen kann viel mehr beschreiben als ihre eigene Bedeutung.

d. **Vollformen:** Einige Vollformen sind fachspezifisch und können nicht erkannt werden. Z. B. beim Terminus „Ingenieurkammer Sachsen-Anhalt“ werden die zwei Termini „Ingenieurkammer“ und „Sachsen-Anhalt“ extrahiert.

Zweisprachige Extraktion:

a **Ein Wort ist gleich ein Satz:** Wegen der spezifischen Merkmale der Textsorte Antrag oder Fomular gibt es bei der Extraktion ähnliche Besonderheiten wie bei Softwaretext (siehe Abbildung 6.1).

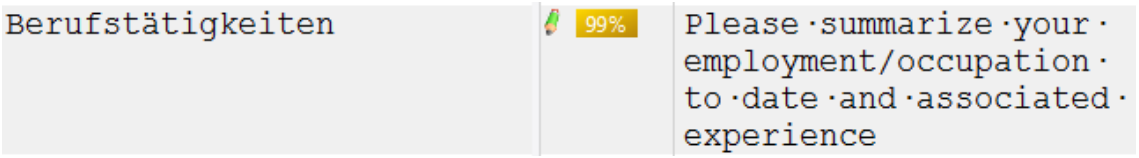


Abbildung 6.1 ein Wort ist gleich ein Satz aus Trados Studio 2014

b **Mehrwortbenennungen mit Ellipse**: Mehrwortbenennungen mit Ellipse sind nicht immer bemerkbar. Z. B. der Begriff „berufliche Ausbildung bzw. Tätigkeit“, der bei der Extraktion als zwei gesonderte Begriffe „berufliche Ausbildung“ und „berufliche Tätigkeit“ extrahiert wird. Aber es wird auch einfach als „berufliche Ausbildung“ und „Tätigkeit“ bezeichnet oder sogar nicht extrahiert. Die folgende Abbildung 6.2 zeigt ein Beispiel an: Zwei Benennungen werden als eine Benennung übersetzt.

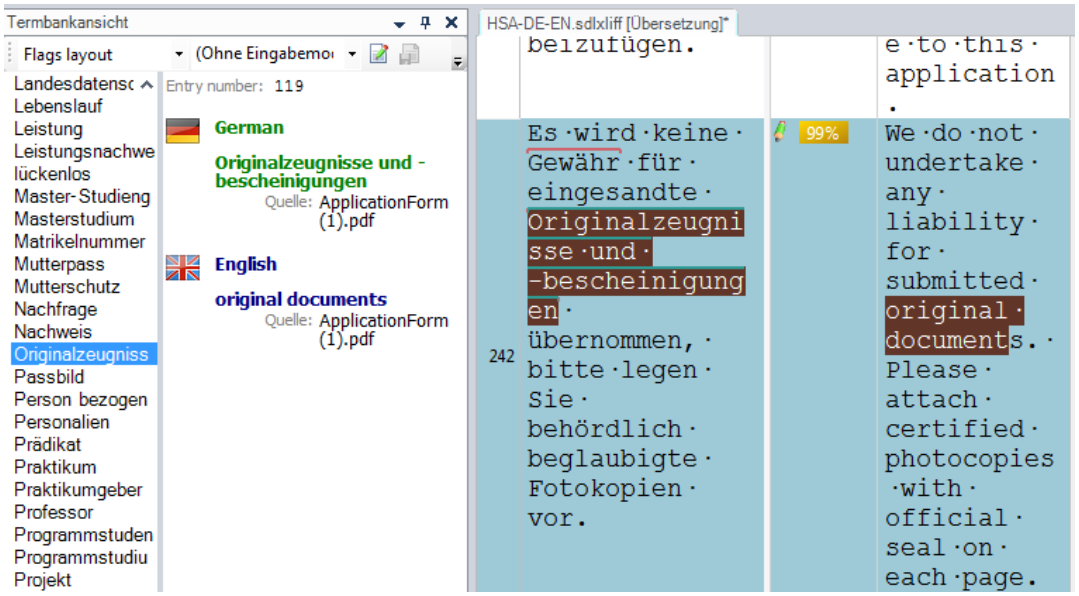


Abbildung 6.2 SDL Trados Studio 2014 - zwei Benennungen werden als eine Benennung übersetzt in Termbankansicht

c **Änderung der Wortart:** Aufgrund der Mehrdeutigkeit zahlreicher Benennungen in Bezug auf die Wortart in Englischen kann ein Substantiv im Englischen mit einen Verb in Deutschen übersetzt werden. Im Satz „Mein bisheriges Studium absolvierte ich an der: (I previuously studied at:)“ wird das Substantiv „Studium“ als ein Verb „study“ übersetzt.

d **Paraphrase**: Es gibt Schwierigkeiten bei der Extraktion von Paraphrasen. Ein Beispiel dafür ist, ob „Studienwunsch und Themenbereich“ im Satz „Begründung des *Studienwunsches und Themenbereichs* für die Masterthesis (explanation for *application to this program and field of interest* for Masterthesis)“ als ein Termnus oder als zwei Termini extrahiert werden soll.

e **Granularität**: Wegen der verschiedenen Übersetzungen erschwert die Extraktion nach dem Kriterium die Granularität (siehe Abbildung 6.3 und 6.4)

Begr	DE	EN
32	amtlich beglaubigte Hochschulzugangsberechtig	authenticated Higher Education Entrance Qualification
33	amtlich beglaubigte Kopie	authenticated copy
33	amtlich beglaubigte Kopie	officially certified copy
33	amtlich beglaubigte Kopie	officially attested copy
34	amtlich beglaubigte Nachweis	officially certified/attested certificates/supporting documents
34	amtlich beglaubigtes Zeugnis	
34	amtliche Beglaubigung	official certification/attestation
35	amtlich beglaubigte Übersetzung	certified translation
35	amtlich beglaubigte Übersetzung	officially certified/ attested translation

Abbildung 6.3 Mehrwortbenennungen in Excel-Tabelle

Begr	DE	EN
32	amtlich	official
33	beglaubigt	certified
34	beglaubigt	attested
34	behördlich beglaubigt	certified
35	amtlich beglaubigt	authenticated
35	amtlich beglaubigt	officially certified
35	amtlich beglaubigt	officially attested
35	amtlich beglaubigt	certified
124	Beglaubigung	certification
124	Beglaubigung	attestation
124	beglaubigte Nachweis	certified/attested c
124	beglaubigte Zeugnis	

Abbildung 6.4 Granularität in Excel-Tabelle

Die Terminologieextraktion ist die Grundlage zur Erstellung eines mehrsprachigen Terminologiebestands. Bei der späteren Verwendung dieser Datenbank muss sie weiter bearbeitet, ergänzt und gepflegt werden.

In der praktischen Arbeit wird die maschinelle Extraktion z. B. mit dem Werkzeug SDL MultiTerm 2014 Extract durchgeführt, um Aufwand zu sparen. Die Qualität der einsprachigen Extraktion (Deutsch) ist gut. Im Gegensatz dazu muss das Verfahren der zweisprachigen Extraktion noch verbessert werden. Weiterhin ist eine bessere Verbindung mit SDL MultiTerm 2014 Desktop erforderlich, da die Zusatzinformationen in SDL MultiTerm 2014 Extract nicht richtig in SDL MultiTerm 2014 Desktop angezeigt werden können. Das PDF-Format ist eines der am häufigsten benutzten Dateiformate bei der Dokumentation. Um den Inhalt und die Struktur der Texte in der PDF-Datei bei der Umwandlung in ein anderes Format zu erhalten, ist die Unterstützung von PDF-Formaten im Extraktionsprogramm wünschenswert.

Literaturverzeichnis

[Arndt 2014] T. Arndt et. al. Modul 5 - Projekt und Prozessmanagement – In: Terminologiearbeit Best Practices. Köln. Deutscher Terminologie Tag e.V. 2014

[Arntz 2014] R. Arntz, H. Picht, K. Schmitz. Einführung in die Terminologiearbeit. 7. vollständig überarbeitete und aktualisierte Auflage. Hilesheim, Zürich, New York. Olms Georg Olms Verlag. 2014

[Bauer 2014] S. C. Bauer et. al. Modul 2 - Grundsätze und Methoden in der Terminologiearbeit – In: Terminologiearbeit Best Practices. Köln. Deutscher Terminologie Tag e.V. 2014

[Childress 2014] M. Childress et. al. Modul 1 - Argumentationshilfen. In: Terminologiearbeit Best Practices. Köln. Deutscher Terminologie Tag e.V. 2014

[DIN 2330] Begriffe und Benennungen: Allgemeine Grundsätze. Berlin. Beuth

[DIN 2342:2004] Begriffe der Terminologielehre. Normenvorlage als Ersatz für DIN 2342-1:1992. Berlin. Beuth

[Drewer 2014] P. Drewer et. al. Modul 3 - Benennungen, in: Terminologiearbeit Best Practices. Köln: Deutscher Terminologie Tag e.V. 2014

[Eckstein 2009] K. Eckstein. Toolgestützte Terminologieextraktion. In: Mayer. F/ Seewalt-Heeg. U (Hrsg.)Terminologiemanagement – Von der Theorie zur Praxis. Schaltungsdienst Lange oHG. Berlin. 2009

[Ferrari 2014] D. Ferrari et. al. Modul 4 - Werkzeuge und Technologien, in: Terminologiearbeit Best Practices. Köln: Deutscher Terminologie Tag e.V. 2014

[Höge 2005] M. Höge, K. M. Ferber. Globale Terminologieverwaltung – eine Herausforderung unserer Zeit. In MultiLingual Computing & Technology. Juli 2005

[IBM 2014] IBM Knowledge Center. Stoppwortlisten. Online-Quelle: <http://www-01.ibm.com/support/knowledgecenter/SSGU8G_11.50.0/com.ibm.excal.doc/excal35.htm%23concepts938187?lang=de>. [07.12.2014]

[Kim 2007] D. Kim. Semantische Analyse und automatische Gewinnung von branchenspezifischem Vokabular für E-Commerce. Dissertation am Centrum für Informations- und Sprachverarbeitung der Ludwig-Maximilians-Universität. München. 2007

[Massion 2009] F. Massion. Terminologiemanagement: Luxus oder Muss. Von der Theorie zur Praxis. In: Mayer. F/ Seewalt-Heeg. U (Hrsg.) Terminologiemanagement – Von der Theorie zur Praxis. Schaltungsdienst Lange oHG. Berlin. 2009

[Massion 2014] F. Massion. Folie: Terminologiemanagement_FH_Anhalt-Massion-2014_Teil_2. 2014

[Müller 2014] K. Müller. Folie: Terminologielehre und Terminologieverwaltung. 2014

[Reineke 2005] D. Reineke, K-D. Schmitz. Einführung in die Softwarelokalisierung. Tübingen. Narr Francke Attempto. 2005

[SDL MultiTerm 2014 Extract] Online-Hilfe von SDL MultiTerm 2014 Extract

[Schmitz 2005 a] K-D. Schmitz. Internationalisierung und Lokalisierung von Software. In: Einführung in die Softwarelokalisierung. Köln. Gunter Narr. 2005

[Schmitz 2005 b] K-D. Schmitz. Terminologieverwaltung für die Softwarelokalisierung. In: Einführung in die Softwarelokalisierung. Köln. Gunter Narr. 2005

[Schmitz 2010] K-D. Schmitz, D. Straub. Erfolgreiches Terminologiemanagement im Unternehmen. Stuttgart. TC and More GmbH. 2010

[Zerfass 2008] A. Zerfass. Terminologiemanagement - Methoden und Programme zur Erstellung, Bearbeitung/ Verwendung und Prüfung von Terminologie. Tekom 2008.

Anhang

1. Dateinamen der Ausgangsmaterialien:

90-DMP_AntragZulassungMasterthesis.pdf

90-DMP_AufgabenstellungMasterthesis.pdf

90-Master_AntragZulassung.pdf

90-Master_Aufgabenstellung.pdf

Anmeldung für andere Prüfungen.pdf

antrag_abschlussarbeit.pdf

antrag_anerkennung_leistungen.pdf

antrag_beurlaubung.pdf

antrag_exmatrikulation.pdf

Antrag_Exmatrikulation.pdf

antrag_master_hsa-KOET.PDF

antrag_master_hsa_BBG.PDF

antrag_stgwechsel.pdf

antrag_studiengangwechsel.pdf

Antrag_Studiengangwechsel.pdf

BachelorAntragVerlaengerung.pdf

Bescheinigung Archiv.pdf

bibliographische_zusammenfassung.pdf

bibo_archivprot.pdf

Formular_Anerkennung Studienleistungen 84-AR PO 2010.pdf

Formular_Anerkennung Studienleistungen 84-DES PO 2012.pdf

Formular_Anerkennung Studienleistungen.pdf

Formular_MAGIS_Beruf_e.pdf

Protokoll Modulpruefung.pdf

Rücktritt_ Prüfung.pdf

Verlängerung_Abschlussarbeit.pdf

zulassung_fachpruefung.pdf

Beurlaubung.pdf

AntragIngenieur.pdf

AntragZulassungBachelorDiplomMasterNeu.pdf

BA_Berufspraktikum_Vertrag.pdf

DesignEignung2013.pdf

Zulant_BA2014-KOET-DE.pdf

Zulant_BA2014-BBG.pdf

Zulant_DU2014.pdf

Zulant_FS2014.pdf

Zulant_Gast2013.pdf

Antrag_auf_Beurlaubung_Application_for_Academic_Leave_of_Absence_Dessau.pdf

Antrag_auf_Exmatrikulation_Application_for_De-Registration_Dessau.pdf

Appl_program_students_2013.pdf

ApplicationForm (1).pdf

Zulant_MA2014-BBG_01.PDF

Zulant_MA2014-KOET.PDF

zulassungsantrag_auslaendbewerb.pdf

application_foreignstudents.pdf

90-Master_AntragZulassung_englisch.pdf

90-Master_Aufgabenstellung_englisch.pdf

2. Bericht zur Umfangsbestimmung für allgemeine Informationen

Gesamtüberblick

Gesamt	Typ	Segmente	Wörter	Zeichen	Prozent	Erkannte Tokens	Tags
Dateien:4	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:6.91	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	649	1388	11126	6.09%	554	298
	Dateiübergreifende Wiederholungen	1198	4697	33191	20.61%	1234	812
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	2300	16704	113233	73.30%	2982	2403
	Gesamt	**4147**	**22789**	**157550**	**100%**	**4770**	**3513**

Datei	Typ	Segmente	Wörter	Zeichen	Prozent	Erkannte Tokens	Tags
⊞ Bernburg.sdlxliff (Zusammengeführte Dateien: 10)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:7.56	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	195	366	3071	9.23%	141	51
	Dateiübergreifende Wiederholungen	0	0	0	0.00%	0	0
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	605	3600	26930	90.77%	961	766
	Gesamt	**800**	**3966**	**30001**	**100%**	**1102**	**817**
⊞ Dessau.sdlxliff (Zusammengeführte Dateien: 16)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:7.18	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	240	562	4130	15.46%	273	202
	Dateiübergreifende Wiederholungen	135	241	1374	6.63%	93	21
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	509	2832	20600	77.91%	817	669
	Gesamt	**884**	**3635**	**26104**	**100%**	**1183**	**892**
⊞ geschützt.sdlxliff (Zusammengeführte Dateien: 3)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:6.07	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	49	121	897	1.76%	9	0
	Dateiübergreifende Wiederholungen	20	27	196	0.39%	2	0
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	627	6710	40562	97.84%	114	4
	Gesamt	**696**	**6858**	**41655**	**100%**	**125**	**4**
⊞ Köthen.sdlxliff (Zusammengeführte Dateien: 18)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:7.18	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	165	339	3028	4.07%	131	45
	Dateiübergreifende Wiederholungen	1043	4429	31621	53.17%	1139	791
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	559	3562	25141	42.76%	1090	964
	Gesamt	**1767**	**8330**	**59790**	**100%**	**2360**	**1800**

3. Bericht zur manuellen Extraktion

Gesamtüberblick

Gesamt	Typ	Segmente	Wörter	Zeichen	Prozent	Erkannte Tokens	Tags
Dateien:5	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:6.89	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	943	2299	17526	11.64%	711	411
	Dateiübergreifende Wiederholungen	363	917	6373	4.64%	272	88
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	2283	16535	112169	83.72%	2976	2407
	Gesamt	**3589**	**19751**	**136068**	**100%**	**3959**	**2906**

Detailansicht

Datei	Typ	Segmente	Wörter	Zeichen	Prozent	Erkannte Tokens	Tags
⊞ DE-einfach.sdlxliff (Zusammengeführte Dateien: 25)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:8.04	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	470	977	8067	24.65%	403	201
	Dateiübergreifende Wiederholungen	0	0	0	0.00%	0	0
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	668	2986	23807	75.35%	1077	897
	Gesamt	**1138**	**3963**	**31874**	**100%**	**1480**	**1098**
⊞ DE-EN-einfach.sdlxliff (Zusammengeführte Dateien: 2)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:6.80	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	10	31	223	7.54%	20	20
	Dateiübergreifende Wiederholungen	16	34	247	8.27%	25	24
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	70	346	2324	84.18%	189	163
	Gesamt	**96**	**411**	**2794**	**100%**	**234**	**207**

Datei	Typ	Segmente	Wörter	Zeichen	Prozent	Erkannte Tokens	Tags
⊞ DE-EN-Kontext.sdlxliff (Zusammengeführte Dateien: 4)	**PerfectMatch**	0	0	0	0.00%	0	0
Zeichen/Wort:6.25	**Kontext-Match**	0	0	0	0.00%	0	0
	Wiederholungen	116	452	3062	4.51%	76	63
	Dateiübergreifende Wiederholungen	60	140	948	1.40%	45	5
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	942	9422	58534	94.09%	917	694
	Gesamt	**1118**	**10014**	**62544**	**100%**	**1038**	**762**
⊞ DE-Kontext.sdlxliff (Zusammengeführte Dateien: 9)	**PerfectMatch**	0	0	0	0.00%	0	0
Zeichen/Wort:7.30	**Kontext-Match**	0	0	0	0.00%	0	0
	Wiederholungen	335	801	5944	15.57%	207	122
	Dateiübergreifende Wiederholungen	280	722	5038	14.03%	194	52
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	567	3622	26562	70.40%	747	611
	Gesamt	**1182**	**5145**	**37544**	**100%**	**1148**	**785**
⊞ EN.sdlxliff (Zusammengeführte Dateien: 2)	**PerfectMatch**	0	0	0	0.00%	0	0
Zeichen/Wort:6.02	**Kontext-Match**	0	0	0	0.00%	0	0
	Wiederholungen	12	38	230	17.43%	5	5
	Dateiübergreifende Wiederholungen	7	21	140	9.63%	8	7
	100%	0	0	0	0.00%	0	0
	95% - 99%	0	0	0	0.00%	0	0
	85% - 94%	0	0	0	0.00%	0	0
	75% - 84%	0	0	0	0.00%	0	0
	50% - 74%	0	0	0	0.00%	0	0
	Neu	36	159	942	72.94%	46	42
	Gesamt	**55**	**218**	**1312**	**100%**	**59**	**54**

4. Bericht zur maschinellen Extraktion

Gesamtüberblick

Gesamt	Typ	Segmente	Wörter	Zeichen	Prozent	Erkannte Tokens	Tags
Dateien:2	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:7.48	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	1223	3625	28781	22.93%	976	541
	Dateiübergreifende Wiederholungen	585	1300	8525	8.22%	206	0
	100%	0	0	0	0.00%	0	0
	95% - 99%	614	4316	30418	27.30%	309	142
	85% - 94%	37	178	1223	1.13%	98	80
	75% - 84%	38	164	1187	1.04%	54	35
	50% - 74%	16	62	467	0.39%	19	11
	Neu	1102	6162	47639	38.98%	1541	1262
	Gesamt	**3615**	**15807**	**118240**	**100%**	**3203**	**2071**

Detailansicht

Datei	Typ	Segmente	Wörter	Zeichen	Prozent	Erkannte Tokens	Tags
⊞ DE in gemischten Dateien.sdlxliff (Zusammengeführte Dateien: 7)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:7.15	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	387	1208	9145	21.51%	102	0
	Dateiübergreifende Wiederholungen	0	0	0	0.00%	0	0
	100%	0	0	0	0.00%	0	0
	95% - 99%	536	4124	28901	73.42%	149	0
	85% - 94%	5	26	179	0.46%	3	0
	75% - 84%	8	46	382	0.82%	3	0
	50% - 74%	5	33	237	0.59%	3	0
	Neu	49	180	1300	3.20%	12	0
	Gesamt	**990**	**5617**	**40144**	**100%**	**272**	**0**
⊞ DE.sdlxliff (Zusammengeführte Dateien: 37)	PerfectMatch	0	0	0	0.00%	0	0
Zeichen/Wort:7.66	Kontext-Match	0	0	0	0.00%	0	0
	Wiederholungen	836	2417	19636	23.72%	874	541
	Dateiübergreifende Wiederholungen	585	1300	8525	12.76%	206	0
	100%	0	0	0	0.00%	0	0
	95% - 99%	78	192	1517	1.88%	160	142
	85% - 94%	32	152	1044	1.49%	95	80
	75% - 84%	30	118	805	1.16%	51	35
	50% - 74%	11	29	230	0.28%	16	11
	Neu	1053	5982	46339	58.70%	1529	1262
	Gesamt	**2625**	**10190**	**78096**	**100%**	**2931**	**2071**