

MASTERARBEIT

Titel der wissenschaftlichen Arbeit

Die Anwendbarkeit des Tools ProTerm für die zweisprachige Terminologieextraktion

Eine Untersuchung unter besonderer Berücksichtigung der Häufigkeitsanalyse und Verwendung von Stopp-Wort-Listen am Beispiel von AVL- Produktliteratur

Verfasserin

Verena Christina Bleich, Bakk.phil.

angestrebter akademischer Grad

Master of Arts (MA)

Pamhagen, im Juni 2010

Studienkennzahl It. Studienblatt: A 060 345 342

Studienrichtung It. Studienblatt: Übersetzen

Betreuer: Univ.-Prof. Mag. Dr. Gerhard Budin

Danksagung

An dieser Stelle möchte ich Herrn Univ.-Prof. Mag. Dr. Gerhard Budin für die wissenschaftliche Betreuung und seine konstruktiven und fachlichen Hinweise danken.

Ein herzlicher Dank gilt Frau Mag. Irmgard Soukup-Unterweger, MAS MSc für ihre wertvollen und unverzichtbaren Ratschläge. Sie hat nicht nur mein Interesse an Terminologiearbeit geweckt, sondern ist für die Inszenierung dieser Masterarbeit verantwortlich.

Herrn Amtsdirektor Hans-Christian Pilles (Zentraldokumentation der Landesverteidigungsakademie) sei ein ganz besonderer Dank für seine kritischen Anmerkungen und sein persönliches Engagement bei der Umsetzung der technischen Änderungen ausgesprochen.

Für die produktive Zusammenarbeit und die reibungslose Koordination seitens der AVL List GmbH möchte ich Herrn Klaus Baumgartner, MSc sehr danken.

Meinen Kolleginnen Maria Kerschbaumer, Bakk. phil., Michaela König, MA und Katharina Spiegl, MA danke ich für ihre freundschaftliche Unterstützung und das gewissenhafte Lektorieren.

Von ganzem Herzen *Danke* für die liebevolle und geduldige Begleitung meiner Familie, in erster Linie meinen Eltern, Maria & Gerhard Bleich. Sie haben mir das Studium an der Universität Wien ermöglicht und meine Pläne stets voller Zuversicht unterstützt. Von ganzem Herzen *Danke* meinem Freund, Markus Landmann DI. Seine motivierenden Worte erfüllen mich immerzu mit Heiterkeit und Kraft.

"Drum seid nur brav und zeigt euch musterhaft, Laßt Phantasie, mit allen ihren Chören, Vernunft, Verstand, Empfindung, Leidenschaft, Doch, merkt euch wohl! nicht ohne Narrheit hören."

¹ Goethe, Johann Wolfgang. Faust – Der Tragödie erster Teil. Wien: Humboldt-Verlag, 1946.



Diese Masterarbeit richtet sich gleichermaßen an Frauen und Männer. Teilweise wurde nicht gendergerecht formuliert, um leichtere Lesbarkeit zu gewährleisten.



Kurzfassung (Deutsch)

Im Rahmen dieser Masterarbeit wurde die Anwendbarkeit des Terminologieextraktionswerkzeuges ProTerm der Firma ProCom Strasser für die zweisprachige Terminologieextraktion untersucht. Der für die Terminologieextraktion notwendige Korpus wurde von der AVL LIST GmbH zur Verfügung gestellt. Dabei handelte es sich ausschließlich um Fachtexte der technischen Dokumentation. Die Aufgabenstellung bestand darin, englische und deutsche Termpaare und adäquate Datenelemente für die Datenkategorien (Definition, Explikation, Kollokation, Kontext) aus dem Dokumentationsmaterial zu extrahieren. Das Terminologieextraktionsverfahren kann als toolgestützt bezeichnet und dem hybriden Extraktionsverfahren zugeordnet werden. Es wurden das statistische und das manuelle Extraktionsverfahren kombiniert. Das Berücksichtigen von Stopp-Wort-Listen der Allgemeinsprache und von Stopp-Wort-Listen, die eigens für diese Arbeit erstellt und bearbeitet wurden, sowie das Koordinieren der individuell kombinierbaren statistischen Parameter trugen dazu bei, Termkandidaten rasch zu identifizieren. ProTerm bietet die Möglichkeit, eine breite Palette an Formaten und unterschiedlichen Zeichensätzen einzulesen, es kann einen großen Umfang an Dokumenten in kurzer Zeit einlesen und ermöglicht es dem Terminologen, während jeder Phase des Extraktionsprozesses in die Originaltextansicht zu wechseln. Jeder Termkandidat und jede Datenkategorie, die der TermBank hinzugefügt werden, werden automatisch mit der dazugehörigen Quelle extrahiert. Mithilfe der Trunkierungsfunktion kann einheitlich Terminologie extrahiert werden. Das Auffinden der zielsprachlichen Äquivalente obliegt allerdings der Kompetenz des Terminologen. Er ist nicht nur dafür verantwortlich, einen Termkandidaten mithilfe der Parametereinstellungen und der Verwaltung der Stopp-Wort-Listen zu identifizieren, sondern muss sein zielsprachliches Äquivalent in den Originaltexten ausfindig machen. Die Erkenntnisse dieser Masterarbeit sollen dazu dienen, die zweisprachige Terminologieextraktion mit ProTerm weiterzuentwickeln.



Abstract (English)

This Master's Thesis examines the terminology management tool ProTerm by Pro-Com Strasser for its applicability to bilingual terminology extraction. AVL List GmbH, "the world's largest privately owned and independent company for the development of powertrain systems with internal combustion engines as well as instrumentation and test systems." (AVL-Company 2010) provided the corpus for the terminology extraction, consisting of technical documentations only. The main task was to extract bilingual term candidates and adequate data for data categories (definition, explication, collocation, and context). The terminology extraction technique applied was semi-automatic and can be assigned to the hybrid approach, combining manual and statistical terminology extraction techniques. Considering stop lists with general language and stop lists especially created for and adapted during this Thesis as well as the coordination of the statistical parameters ad libitum facilitated the term recognition within a short time. Moreover ProTerm succeeds in importing various formats (Microsoft Office files, .pdf, .txt, .html, and .xml) and character sets (ISO 8859-1 Western Europe and UTF-8) as well accessing the source text at any level. Every term candidate (and every data category) added to the TermBase is extracted automatically together with its source. Truncation allows keeping consistency of terminology. The terminologist has to identify term candidates and their corresponding equivalents in the target texts by combining parameter settings with the administration of stop lists. The results of this Master's Thesis shall contribute to further develop bilingual terminology extraction with ProTerm.



1. Inhaltsverzeichnis

1.	Inh	altsv	/erzeichnis	1
2.	Ein	leitu	ng	5
3.	Zie	le de	er Arbeit	7
4.	Tei	rmine	ologieextraktion	9
4	.1	Tex	ttyp, Textsorte	9
4	.2	Wa	s ist Terminologieextraktion?	10
4	.3	Ter	minologieextraktionsverfahren	10
	4.3	.1	Manuelle Extraktion	10
	4.3	.2	Toolgestützte Terminologieextraktion	11
	4.3	.3	Konkordanzwerkzeuge	11
	4.3	.4	Statistische Extraktionsverfahren	12
	4.3	.5	Linguistische Extraktionsverfahren	13
	4.3	.6	Hybride Extraktionsverfahren	13
	4.3	.7	Evaluierungskriterien für Terminologieextraktionstools	14
5.	Ko	oper	ationspartner	15
5	5.1	Pro	Term	15
	5.1	.1	ProCom Strasser & DocuMatrix	15
	5.1	.2	Zentraldokumentation der Landesverteidigungsakademie	15
	5.1	.3	Vorgaben ProTerm- Kooperationspartner	16
5	5.2	AVL	LIST GmbH	16
	5.2	.1	Vorgaben AVL	17
6.	Exl	kurs:	Datenkategorien	19
6	5.1	Def	inition	19
6	5.2	Exp	olikation	20
6	5.3	Koll	lokation	20
6	5.4	Kor	ntext	21

7.	Pr	oTer	m	23
	7.1	Vor	bereitende Maßnahmen	. 23
	7.2	Pro	zess Terminologieextraktion mit ProTerm	. 24
	7.2	2.1	Vorbereiten der Texte	26
	7.2	2.2	ProTerm starten	26
	7.2	2.3	Projekt und Filter anlegen/auswählen	27
	7.2	2.4	InTerm Einlesen der Dokumente	31
	7.2	2.5	NewTerm Terminologieextraktion	35
	7.3	Hin	zufügen von Termkandidaten	. 52
	7.4	lde	ntifizieren von Termkandidaten	. 56
	7.5	Rev	vision	57
	7.6	Ter	mBank	. 57
	7.0	6.1	Termbankinhalt	60
	7.0	6.2	Begriffsebene	60
	7.0	6.3	Sprachebene	61
	7.0	6.4	Termebene	62
	7.7	Exp	oort aus ProTerm	63
	7.8	Erg	ebnis der Terminologieextraktion	67
	7.9	Val	idierung der Termkandidaten	. 68
	7.10	S	topp-Wort-Listen	. 68
	7.	10.1	Erstellen neuer Stopp-Wort-Listen	70
	7.	10.2	Verwendung der Stopp-Wort-Listen	72
	7.	10.3	Generieren von Stopp-Wörtern während des Auswahlverfahrens	73
	7.	10.4	Ändern der Stopp-Wortlisten	74
	7.11	Α	ndere Methoden	77
	7.	11.1	Ein Filter mit allen Dokumenten einer Dokumentationsgruppe	77
	7.	11.2	pdf-Dokumente kapitelweise einlesen	77
8.	Sc	chlus	sbetrachtung	79

9. Lite	. Literaturverzeichnis		
10. A	nhang	87	
10.1	Benutzeroberfläche ProTerm	87	
10.2	ProTerm-Funktionstasten	92	
10.3	Abkürzungsverzeichnis	93	
10.4	Tabellenverzeichnis	93	
10.5	Abbildungsverzeichnis	94	
10.6	Index	96	
11. C	Curriculum Vitae	97	

2. Einleitung

Zu Beginn soll auf die Bedeutung von Terminologie in der heutigen Zeit hingewiesen werden:

"Terminologie als Gesamtheit der Begriffe eines Fachgebiets ist heute von enormer Bedeutung für eine Reihe von wirtschaftlich interessanten Gebieten. Sie ist das Skelett jeder fachwissenschaftlichen und spezialisierten Kommunikation – so etwa bei wissenschaftlichen Kongressen (…). Sie spielt eine große Rolle in der zunehmenden internationalen Zusammenarbeit und in der Gesetzgebung, wo die Äquivalenzen zweier Begriffe genau geregelt sein müssen" (Haller 2007).

"Die Verwendung einer korrekten Terminologie ist heutzutage in vielen Lebensbereichen von zunehmender Bedeutung. Eine exakte und vollständige Terminologie steigert die Produktivität von Übersetzern und technischen Redakteuren und ist eine Voraussetzung für erfolgreiche Kommunikation. Insbesondere Arbeitsfelder wie technische Dokumentation, Übersetzung und Softwarelokalisierung erfordern eine systematische Terminologiearbeit. Daher wird die Terminologieverwaltung zu einer immer wichtigeren Aktivität bei der Vorbereitung, Bearbeitung und Dokumentation eines Fachwortschatzes" (Zielinski und Safar 2005).

Um Terminologie rascher verfügbar zu machen und effizient nutzen zu können, fügt Lieske (2002) hinzu, dass "(…) Industrieunternehmen mit großem Bedarf an Terminologie¹ sich daher für Werkzeuge und Dienstleistungen für die Term-Extraktion interessieren".

Im ersten Teil dieser Masterarbeit werden ihre Ziele und die Herangehensweise an die Aufgabenstellung erläutert. Im Anschluss werden wissenschaftliche Ansätze zur Terminologieextraktion (TE)² vorgestellt, wobei zunächst darauf geachtet wurde, dass diese für die Arbeit mit dem Terminologieextraktionstool (TET)³ ProTerm von Relevanz sind; um das Kapitel zu vervollständigen werden danach auch andere Ansätze kurz vorgestellt.

¹ Wie im vorliegenden Fall die AVL LIST GmbH (siehe 5.2.1).

² "Insbesondere im Englischen existieren viele Synonyme für das Konzept der Term-Extraktion (z. B. terminology extraction, terminology mining, automatic terminology detection oder terminology identification). (Lieske 2002) Witschel (2005) wählt die Schreibweise Terminologie-Extraktion. Nach Lieske (2002), Zielinski und Safar (2005), Mügge (2007) und Eckstein (2009) wird im Rahmen dieser Arbeit der Terminus Terminologieextraktion mit der Kurzform (TE) verwendet.

In dieser Masterarbeit wird die Benennung *Terminologieextraktionstool* (*TET*) verwende, vgl. Zielinski und Safar (2005), Zerfaß (2006) spricht von *Termextraktionsprogrammen*, Eckstein (2009) von *Terminologieextraktionsprogrammen* (*TEP*) und Lieske (2002) von *Term-Extraktions-Werkzeugen* (*TEW*).

Im anschließenden Teil werden die Kooperationspartner dieser Masterarbeit, sowie ihre Anforderungen an das Projekt "Terminologieextraktion mit ProTerm" präsentiert. Ein kurzer Exkurs gibt im Anschluss einen Überblick über die extrahierten Datenelemente für Datenkategorien. Der folgende Teil stellt das TET ProTerm vor und zeigt, wie die im Rahmen dieser Arbeit die zweisprachige TE stattgefunden hat. Zum Abschluss wird die Arbeit mit ProTerm kritisch analysiert.

3. Ziele der Arbeit

Das Ziel dieser Arbeit besteht darin, das TET ProTerm für die zweisprachige TE zu testen und den Terminologiebestand der AVL LIST GmbH zu erweitern. Zu diesem Zweck hat die AVL LIST GmbH das notwendige Datenmaterial zur Verfügung gestellt. In dieser Arbeit soll festgestellt werden, ob TE mit ProTerm möglich ist und wie mit ProTerm zweisprachig Terminologie extrahiert werden kann. Besondere Beachtung bei der TE mit ProTerm wird der Verwaltung von Stopp-Wort-Listen (StW-Listen) und der Häufigkeitsanalyse geschenkt. Durch die Koordination dieser beiden Schritte soll nämlich das manuelle Zutun und somit der Einsatz humaner Ressourcen im Zuge der TE so gering wie möglich gehalten werden. Die im Zuge der TE gewonnenen Termini sollen dazu dienen, den Terminologiebestand der AVL zu erweitern. Aus den zur Verfügung gestellten Dokumentationen sollen nicht nur Termini extrahiert werden, sondern - soweit vorhanden und als relevant erachtet - auch Datenelemente für andere Datenkategorien. Ein zweitrangiges Ziel dieser Arbeit besteht darin, die Entwickler von ProTerm dabei zu unterstützen, das Tool weiter zu entwickeln und seine Effizienz bei der Terminologiearbeit zu erhöhen. Es wird darauf abgezielt, so viele Schritte wie möglich während des Extraktionsprozesses zu automatisieren, also mithilfe des TETs durchzuführen.

4. Terminologieextraktion

4.1 Texttyp, Textsorte

In diesem Kapitel werden die Texte, die im Zuge dieser Masterarbeit bearbeitet wurden, analysiert. Es handelt sich um Texte, "die primär Informationen vermitteln" (Kadric, et al. 2005:78), und daher können sie dem informativen Texttyp zugeordnet werden.

"Die Bezeichnung dieser Textsorte ist im Deutschen etwas problematisch. Im Englischen trifft die hyperonymische Bezeichnung *manuals* den gemeinten Begriff recht gut (...) Gelegentlich verwendet man (...) auch im Deutschen den Ausdruck *Manual* als Lehnwort (mit englischer Aussprache), selten auch phonetisch und hinsichtlich Deklination ans Deutsche assimiliert als Manual (pl: Manuale). Gemeint sind damit jene Teile einer Produktdokumentation, in denen der Benutzer eines Produkts mit dem Produkt und dessen Gebrauch, Bedienung, Betrieb, Instandhaltung und/oder Instandsetzung vertraut gemacht werden soll. (...) Im Gegensatz zum en. Ausdruck *manual*, der von einem einzelnen Blatt über ein geklammertes Heft bis zu mehrbändigen Büchern alles abdeckt, kann der dt. Ausdruck *Handbuch* nur auf solche Dokumentationen bezogen werden, die tatsächlich die Merkmale eines Buches (Bindung mit Rücken, auch Ringbücher) aufweisen" (Schmitt 2003).

Bei den von der AVL zur Verfügung gestellten Texten handelt es sich um Fachtexte.

"Ein Wesensmerkmal von Fachtexten ist deren sprachliche Spezialisierung. Diese Spezialisierung kommt vor allem in der Verwendung von fachspezifischen Benennungen zum Ausdruck, d. h., in der Fachsprache werden Benennungen verwendet, die entweder in der Gemeinsprache überhaupt nicht verwendet werden (z. B. Benutzeroberfläche in der Informatik) oder in der Fachsprache für einen anderen Begriff stehen (z. B. Mutter in der Mechanik) als in der Gemeinsprache. Deshalb können Fachtexte nur dann sachlich richtig von einer Sprache in eine andere übertragen werden, wenn bei der Übersetzung die entsprechende multilinguale Terminologie zur Verfügung steht" (Mügge 2007).

4.2 Was ist Terminologieextraktion?

Zerfaß (2006) stellt fest, dass "Extraktion von Terminologie […] eine sehr subjektive Angelegenheit" ist.

"Terminologieextraktion kann als Prozess zur Identifizierung von Termkandidaten (TK) in einem gegebenen Text definiert werden und ist terminologisch von Termerkennung zu unterscheiden. Termerkennung bezeichnet den Prozess des Vergleichs von Termkandidatenlisten (die Ausgabe von TE) mit einer bestehenden Termdatenbank (TDB) mit dem Ziel, bekannte von unbekannten Termini zu unterscheiden" (Zielinski und Safar 2005).

Zielinski und Safar (2005) unterscheiden monolinguale und bilinguale TE:

"Monolinguale TE wird beim Übersetzungsprozess normalerweise vor Begin (sic!) des Übersetzens von Ausgangs- oder Referenztexten angewendet. Das Ziel ist die Erkennung der relevanten Terminologie eines zu übersetzenden Texts oder – im Fall der reinen Terminologiearbeit – die Erkennung der Termini eines gewissen Fachgebiets. Bilinguale TE wird hingegen hauptsächlich auf übersetzte Texte (paralleler Korpora oder Translation Memories) angewendet. Das Hauptziel dabei ist die Erkennung potentieller Äquivalente in beiden Sprachen (Thurmair, 2003). In beiden Fällen können die gewonnenen TK mit bereits existierenden Termdatenbanken verglichen werden, um bekannte Termini von unbekannten zu differenzieren (vgl. Saß 2004)."

4.3 Terminologieextraktionsverfahren

Zielinski und Safar (2005) sowie Witschel (2005) klassifizieren die Ansätze zur TE als linguistisch, statistisch oder hybrid. Zerfaß (2006) unterscheidet vier Extraktionsverfahren: manuelle Extraktion, Konkordanzprogramme, Statistische und Linguistische Extraktionsverfahren. Mügge (2007) unterteilt die Verfahren in Manuelle Terminologieextraktion, Anwendungen mit Indexwerkzeugen oder Komplexen Konkordanzwerkzeugen. Eckstein (2009) gliedert die Methoden der toolgestützten Terminologieextraktion in automatische, halbautomatische und manuelle TE. Im Folgenden werden die einzelnen Extraktionsverfahren im Detail vorgestellt.

4.3.1 Manuelle Extraktion

Bei der manuellen Extraktion wird der Text vom Übersetzer oder Terminologen gelesen und verstanden und "dieser kann aufgrund seines Vorwissens im

Fachgebiet oder der Zielsetzung für die Extraktion (...) entscheiden, welcher Terminus oder welches Termpaar in die Liste aufgenommen wird" (Zerfaß 2006).

Mügge (2007) fügt hinzu, dass während der Lektüre "Terminologiekandidaten ggf. mithilfe geeigneter Makros in eine Extraktionsliste eingetragen" werden. Er merkt ferner an, dass

"[d]ieses Verfahren die einzige allgemein bekannte Methode zur Terminologieextraktion ist, die sich jedoch ungeachtet ihres geringen Komplexitätsgrades keiner großen Beliebtheit erfreut, da die manuelle Extraktion insbesondere bei großen Übersetzungsprojekten ausgesprochen personal, zeit (sic!) und kostenintensiv ist" (Mügge 2007).

4.3.2 Toolgestützte Terminologieextraktion

"Neben der manuellen Terminologieextraktion kann sich der Terminologe auch von spezieller Software zur Terminologieextraktion unterstützen lassen (…) Genau wie die Systeme zur automatischen Terminologieextraktion filtern halbautomatische TEP⁴ die Termkandidaten auf der Basis linguistischer, statistischer oder kombinierter Verfahren aus einem Text (…) Bei der halbautomatischen TE ist anschließend die Prüfung und Entscheidung durch den Terminologen notwendig. Dieser beurteilt, ob ein Termkandidat tatsächlich ein Terminus ist und in die Terminologiesammlung aufgenommen wird. Bei der automatischen TE entfällt dieser Schritt, da die Software selbst eine Überprüfung und Gewichtung der TK vornimmt" (Eckstein 2009).

4.3.3 Konkordanzwerkzeuge

"Mithilfe von Konkordanzprogrammen oder Konkordanzfunktionen in Translation-Memory-Systemen kann eine Liste aller Wörter erstellt werden. Je nach Einstellungen können hier Einwort- und/oder Mehrwort-Termini aufgelistet werden. Diese Listen sind meist sehr umfangreich, allerdings bieten sie den Vorteil, dass keine Termini vergessen werden" (Zerfaß 2006).

"Stehen Paralleltexte, d. h. Originaltext und Übersetzung, in maschinenlesbarer Form zur Verfügung, werden in einem ersten Schritt die einzelnen Sätze/Segmente im Quelltext ihrer jeweiligen Entsprechung in der Übersetzung zugeordnet. In einem zweiten Schritt wird dann mithilfe linguistischer und/oder

-

⁴ Eckstein (2009) verwendet die Abkürzung *TEP* für *Terminologieextraktionsprogramme*. In dieser Arbeit wird dafür der Terminus *Terminologieextraktionstool* (*TET*) verwendet.

statistischer Methoden, ggf. unter Einsatz von Wörterbüchern, ein zweisprachiges Glossar der in diesen Texten verwendeten fachsprachlichen Benennungen erzeugt" (Mügge 2007).

4.3.4 Statistische Extraktionsverfahren

Zielinski und Safar (2005) stellen folgende Überlegungen zu statistischen Extraktionsverfahren in den Raum:

"Statistische Ansätze basieren auf der Annahme, dass die Wiederholung gewisser lexikalischer Einheiten oder morphosyntaktischer Konstruktionen charakteristisch für Fachtexte ist. Durch die Anwendung verschiedener statistischer Methoden (...) filtern statistisch basierte TETs Wörter und Phrasen aus einem Text heraus, die mit einer Häufigkeit im Text vorkommen, die über einem gegebenen Schwellenwert liegt. Termbasierte statistische Methoden berechnen die Struktur eines TK beispielsweise anhand der N-Grammstruktur⁵. Oft wird die Struktur existierender Termini mit der von Wörtern oder Phrasen eines Korpus verglichen, um TK mit ähnlichen N-grammstrukturen (sic!) herauszufiltern [...]. Eine weitere verbreitete Methode stützt sich auf die Annahme, dass Termini in Fachtexten häufiger als in allgemeinsprachlichen Texten vorkommen, und vergleicht die Häufigkeiten von Wörtern und Phrasen in einem Fachtext mit den Häufigkeiten dieser Einheiten in einem allgemeinsprachlichen Text."

Mügge bedient sich bei statistischen Extraktionsverfahren der Indexwerkzeuge. Dabei werden "Listen sämtlicher in einem Text verwendeten Wörter, die ggf. mit einem bestehenden Wörterbuch oder so genannten Stopplisten abgeglichen" (Mügge 2007).

Zerfaß (2006) beschreibt statistische Extraktionsverfahren folgendermaßen:

"Bei einsprachigen Dokumenten werden die Termini anhand von relativer Häufigkeit aus dem Text extrahiert. (...) Bei zweisprachigen Dokumenten (z. B. Translation-Memory-Dateien oder bilingualen Dateien aus der Übersetzung kommt noch die Suche nach der passenden Übersetzung des Terminus in der Zielsprache dazu. Auch hier wird in statistischen Systemen nach der Häufigkeit entschieden, mit der ein Terminus in der Übersetzung mit dem Terminus in der Ausgangssprache korrespondiert. Da für dieses Verfahren kein morphologisches Hintergrundwissen nötig ist, kann man mit einem statistisch arbeitenden System alle Sprachen bearbeiten."

⁵ N-Grame sind eine Folge benachbarte Elemente (Buchstaben oder Wörter) (sic!).

4.3.5 Linguistische Extraktionsverfahren

"Für dieses Verfahren benötigt das System umfangreiches Wissen über die Sprache, aus der extrahiert wird. Termini werden nicht aufgrund der Häufigkeit extrahiert, sondern das System "versteht", wo in einem Satz z. B. das Subjekt oder das Objekt steht, mit welchen anderen Wörtern es häufig zusammen vorkommt. Als extrahierter Terminus wird es in der Regel auf seine Grundform zurückgeführt und als Termkandidat markiert. Die linguistische Analyse der Sprache erfordert ein großes Wörterbuch der Sprache sowie ein Regelwerk. Daher sind Extraktionsprogramme, die linguistisch arbeiten auf die Sprachen beschränkt, für die sie entwickelt wurden" (Zerfaß 2006).

Zielinski und Safar (2005) führen aus, dass

"(I)inguistisch basierte TETs Termini anhand ihrer morphologischen oder syntaktischen Struktur erkennen. Dazu werden in einem ersten Schritt Texte von morphologischen Analyseprogrammen, Wortarten-Taggern und Parsern mit linguistischer Information annotiert. Dann werden die TK mit einer bestimmten Tagstruktur aus dem annotierten Text mit Hilfe von Methoden der Mustererkennung (Pattern matching) herausgefiltert. Bei den termbasierten Methoden werden TK nach ihrer inneren Struktur gefiltert, z. B. nach ihrer morphologischen Struktur (beispielsweise "Zylinderabschaltung" ds=zylinder#ab_\$ schalten~ung). Bei kontextbasierten Methoden werden TK durch die Analyse der morphosyntaktischen Struktur eines Wortes oder einer Phrase erkannt, d. h. durch die Filterung einer Wortarten-Abfolge wie NP= Nomen + Nomen (e.g. printer menu). Eine weitere Technik beruht auf der Filterung von TK durch die Erkennung von häufig verwendeten Textstrukturen wie Definitionen und erläuternden Kontexten, z. B. "X wird als ... bezeichnet" oder "X besteht aus... (vgl. Pearson 1998, Saß 2004)."

4.3.6 Hybride Extraktionsverfahren

"TETs, die rein linguistisch oder rein statistisch arbeiten, scheitern bei der Lösung vieler typischer Probleme der TE. (...) Da diese Probleme zum Teil sehr unterschiedlicher Natur sind, scheint allein eine Kombination beider Ansätze die Entwicklung effizienter TETs zu ermöglichen. Deshalb wird der so genannte hybride Ansatz wegen seiner "unerforschten" Möglichkeiten von mehreren Autoren als die einzige viel versprechende Methode angesehen" (Zielinski und Safar 2005).

Bei den hybriden Extraktionsverfahren werden also zwei oder mehr der oben angeführten Extraktionsverfahren kombiniert. Das Terminologieextraktionsverfahren, das im Zuge dieser Masterarbeit zur Anwendung kam, kann als toolgestützt bezeichnet

und dem hybriden Extraktionsverfahren zugeordnet werden. Es wurden das statistische und das manuelle Extraktionsverfahren kombiniert. Der detaillierte Ablauf des Extraktionsverfahrens mit ProTerm wird in Kapitel 7 vorgestellt.

4.3.7 Evaluierungskriterien für Terminologieextraktionstools

"TETs können anhand verschiedener Kriterien bewertet werden. Neben grundsätzlichen Funktionalitätsparametern wie Auswahl unterstützter Sprachen und Dateiformate, ist die Qualität der extrahierten Termkandidaten das entscheidende Bewertungskriterium [...]. Die Genauigkeit von TETs wird in der Regel mit den Maßen noise und silence sowie recall und precision ausgedrückt. Während noise sich auf das Verhältnis zwischen den abgelehnten und den angenommenen TK bezieht, gibt silence die Anzahl der von einem TET nicht erkannten Termini an. Recall und precision sind zwei Maße, die oft im IR angewendet werden. Das erstere wird als das Verhältnis zwischen der Summe korrekt gewonnener Termini und der Summe der existierten Termini definiert; das letztere als das Verhältnis zwischen korrekt extrahierten Termini und der Summe vorgeschlagener TK (vgl. Zielinski 2002)" (Zielinski und Safar 2005).

Eckstein (2009) stellt fest, dass es "für die Evaluierung und den Vergleich eines Terminologieextraktionsprogramms bislang kein standardisiertes Modell gibt". Sie stellt folgende Bewertungskriterien vor:

"Noise bezieht sich (...) auf das Verhältnis zwischen relevanten und irrelevanten Termkandidaten, gilt also als Maß für "ungewollte" extrahierte Termkandidaten, die anschließend manuell vom Terminologen aus der Ergebnisliste gelöscht werden. Als Silence werden die Termkandidaten bezeichnet, die bei der Extraktion unentdeckt bleiben und ebenfalls manuelle Nacharbeit (Nacherfassung) erfordern. (...) Während der Recall Auskunft darüber gibt, wie viele relevante Termkandidaten im Verhältnis zu Gesamtzahl relevanter Termkandidaten innerhalb des Textmaterials vom Extraktionsprogramm gefunden werden, gibt Precision an, wie viele vom Programm ausgegebene Kandidaten wirklich relevant sind, und hat somit wiederum Einfluss auf den Nachbearbeitungsaufwand durch den Terminologen. (...) Als weitere Evaluierungskriterien für TEP werden in der Literatur technische Aspekte, die Bedienoberfläche sowie Benutzerfreundlichkeit allgemein, unterstützte Formate beim Import und Export und andere Möglichkeiten des Datenaustauschs, die Unterstützung von Sprachen und Mehrsprachigkeit, Parametrisierbarkeit, Methoden im Validierungsprozess sowie ökonomische Aspekte genannt" (Eckstein 2009).

5. Kooperationspartner

5.1 ProTerm

5.1.1 ProCom Strasser & DocuMatrix

"ProCom-Strasser versteht sich als umfassende (sic!) Partner für den effizienten Umgang mit Content. Die Palette beinhaltet das zielgenaue Beschaffen von relevanten Informationen in unterschiedlichen internen und externen Datenquellen, die Verwaltung von Informationen auch in Terabyte-Mengen, den Aufbau, die Pflege und den Einsatz von Thesauri und semantischen Netzwerken" (Semantic Web Company 2010).

ProTerm ist das "Werkzeug für (…) Terminologie-Verwaltung zur Erstellung und Pflege komplexer Thesauri und Semantischer Netze" (ProCom-Strasser 2009).

Die DocuMatrix Output- und Informationstechnologie Consulting GmbH ist in "Beratung (...), Produkt- und Lösungsverkauf bis hin zur Implementierung und Wartung von Lösungen, die im Web und Multichannel output Bereich angesiedelt sind" (DocuMatrix 2007), tätig. Ihre Aktivitäten konzentrieren sich auf "Firmenweite Outputlösungen: hochvolumiger (batchorientierter) Output, transaktionaler Output, Preview, Browserbasierende Administration und Überwachung sowie Webapplikationen mit Client-/Server ähnlicher Charakteristik" (DocuMatrix 2007).

DocuMatrix hat gemeinsam mit ProCom Strasser ProTerm entwickelt, "um das systematische Arbeiten mit Terminologie zu erleichtern" (DocuMatrix 2007).

Weitere Informationen über ProCom Strasser und DocuMatrix sind auf folgenden Websiten zu finden: www.procom-strasser.com und www.documatrix.com.

5.1.2 Zentraldokumentation der Landesverteidigungsakademie

Im Rahmen einer Ablöse einer Suchmaschine für das Österreichische Bundesheer wurde für die Unterstützung einer semantischen Suche die Entwicklung von Pro-Term von Anwenderseite unterstützt. Amtsdirektor Hans Christian Pilles von der Zentraldokumentation der Landesverteidigungsakademie (Bundesministerium für Landesverteidigung) war der Hauptansprechpartner für alle ProTerm-Anliegen und technischen Fragen.

"Die Zentraldokumentation der Landesverteidigungsakademie ist die interne militärische Fachinformationsstelle für das Österreichische Bundesheer. Ihre Aufgabe ist es, aus eigenen Datenbanken, dem Internet, Zeitungen, Zeit-

schriften und sonstigen Druckwerken laufend relevante Fachinformationen auszuwerten, zu dokumentieren und den internen Bedarfsträgern zur Verfügung zu stellen" (Österreichs Bundesheer 2010).

5.1.3 Vorgaben ProTerm- Kooperationspartner

Die wichtigsten Vorgaben der ProTerm-Kooperationspartner für diese Masterarbeit bestanden darin, ProTerm für die zweisprachige Terminologieextraktion zu testen und Anregungen zu liefern, wie mit ProTerm effizient zweisprachig Terminologie extrahiert werden kann. Ein weiterer Beweggrund für die Bereitstellung von ProTerm war es, etwaige Schwachstellen des Terminologieverwaltungswerkzeuges zu eruieren und zu dokumentieren.

5.2 AVL LIST GmbH

"Die Firma AVL LIST GmbH⁶ gilt als Paradebeispiel für ein international tätiges, exportorientiertes Unternehmen der österreichischen Industrie. (…) Die Homepage von AVL gibt unter der Adresse www.avl.com über Geschichte und Unternehmensbereiche umfassend Auskunft" (Soukup- Unterweger 2002).

"AVL ist das weltweit größte private und unabhängige Unternehmen für die Entwicklung von Antriebssystemen mit Verbrennungsmotoren und Mess- und Prüftechnik. AVL ist in folgenden Unternehmensbereichen tätig: Entwicklung von Antriebssystemen: AVL entwickelt und verbessert alle Arten von Antriebssystemen als kompetenter Partner der Motoren- und Fahrzeugindustrie. Simulation: Die für die Entwicklungsarbeiten notwendigen Simulationsmethoden werden ebenfalls von AVL entwickelt und vermarktet. Motorenmesstechnik und Testsysteme: Die Produkte dieses Bereiches umfassen alle Geräte und Anlagen, die für das Testen von Motoren und Fahrzeugen erforderlich sind" (AVL-Unternehmen 2010).

"Derzeit werden für diese Produkte ca. 32.000 Seiten Kundendokumentation betreut. Die Dokumentation wird zunächst auf Deutsch erstellt und danach ins Englische übersetzt. Weitere Übersetzungen erfolgen je nach Bedarf in den Tochterunternehmen. Die Übersetzungen werden mit dem Translation-Memory-System Transit der Firma STAR erstellt (…)" (Gasser 2004).

⁶ Im Folgenden wird das Unternehmen ,AVL LIST GmbH' kurz als ,AVL' geführt.

5.2.1 Vorgaben AVL

AVL hat aus vier Dokumentationsgruppen (siehe Tab. 1) - CAMEO, EMCON, Indiziertechnik und SANTORIN – neun pdf-Dokumente von insgesamt 2844 Seiten in deutscher und englischer Sprache in elektronischer Form zur Verfügung gestellt. Die Vorgaben von AVL lauteten, möglichst viele Fachtermini aus allen in den Dokumentationen vorhandenen Fachgebieten und adäguate Daten für zusätzliche Datenelemente für Datenkategorien (Definitionen, Kontexte, Explikationen und Kollokationen – siehe Kapitel 6) zu extrahieren, sowie auf etwaige synonyme Verwendungen aufmerksam zu machen. Die AVL-Terminologiearbeit sieht keine Verwendung von Synonymen in ihren Dokumentationen vor. In der AVL-Terminologiedatenbank wird daher auf Negativbenennungen, also Benennungen, die in AVL-Dokumentationen nicht verwendet werden sollen, hingewiesen. Da es mitunter vorgekommen ist, dass dennoch Synonyme verwendet wurden, sollte diese Arbeit dazu beitragen, sie aufzuzeigen. Das Qualifzieren der Synonyme als akzeptierte Benennungen beziehungsweise Negativbenennungen, obliegt den AVL-Mitarbeitern im Zuge der Validierungsphase (siehe Kapitel 7.9) und ist nicht Teil dieser Masterarbeit.

Tab. 1 Übersicht Extraktionsmaterial

			de & en
Dokumentationsgruppe	de (Seiten)	en (Seiten)	(Seiten)
САМЕО	378	376	754
EMCON			
Systemhandbuch	310	310	620
Prüfstandkupplung	34	34	68
Kalibrierung_Drehmoment_Messflansch_Messwelle	62	62	124
Reifenschlupfsimulation	42	42	84
EMCON_gesamt	448	448	896
Indiziertechnik	212	208	420
SANTORIN			
Benutzerhandbuch	158	160	318
Daten Manager	194	192	386
Security Manager	38	32	70
SANTORIN_gesamt	390	384	774
GESAMT	1428	1416	2844

6. Exkurs: Datenkategorien

In Kapitel 5.2.1 wurde bereits erwähnt, dass es eine der AVL-Vorgaben war, Datenelemente für Datenkategorien aus den zur Verfügung gestellten Dokumentationen zu extrahieren.

In diesem Zusammenhang soll auf die Erkenntnisse der Onlineumfrage von Zielinski und Safar (2005) hingewiesen werden: "(...) Übersetzer, Dolmetscher und Terminologen sind nicht nur daran interessiert, Termini und deren zielsprachliche Entsprechungen zu extrahieren und zu speichern, sondern wollen auch zusätzliche Informationen sammeln wie z. B. Kontexte, Definitionen oder andere semantische Informationen (z. B. semantische Relationen). Deshalb wäre ein Schritt der Hersteller von TETs in die Richtung derartiger Entwicklungen sehr willkommen."

ProTerm ermöglicht es, Datenelemente für Datenkategorien, die in den eingelesenen Texten vorhanden sind, ohne großen Zeitaufwand zu extrahieren (siehe Kapitel 7.2.4). Im Vorfeld wurde mit der AVL vereinbart, Datenelemente für die Datenkategorien *Kontext*, *Definition*, *Explikation* und *Kollokation* je nach Vorkommen in den Dokumentationen zu extrahieren. Im Folgenden wird näher auf die extrahierten Datenelemente für die Datenkategorien näher eingegangen.

6.1 Definition

Schmitz (2003) erläutert die Bedeutung von Definition wie folgt: "Aussage, die einen Begriff beschreibt und die Abgrenzung von anderen Begriffen innerhalb eines Begriffssystems ermöglicht".

"Für Terminologielehre und Terminologiearbeit sind Definitionen ganz besonders wichtig, denn hier stehen die Begriffe im Mittelpunkt, und diese müssen mit sprachlichen Mitteln eingegrenzt bzw. beschrieben werden. Entsprechend lautet die Definition von "Definition" in DIN 2342: Begriffsbestimmung mit sprachlichen Mitteln. (1992:2) Zur Funktion von Definitionen sagt DIN 2330 (1993:6) folgendes: Beim Definieren wird ein Begriff mit Hilfe des Bezugs auf andere Begriffe innerhalb eines Begriffssystems festgelegt und beschrieben und damit gegenüber anderen Begriffen abgegrenzt. Die Definition bildet die Grundlage für die Zuordnung einer Benennung zu einem Begriff; ohne sie ist es nicht möglich, einem Begriff eine geeignete Benennung zuzuordnen. Zusätzliche Informationen enthält die Definition von Dahlberg (1981:17): A definition is the equivalence between a definiendum ("what is to be defined?") and a definiens ("how is something to be defined?") for the purpose of delimiting the understanding of the definiendum in any communication case. (…) Die Definiti-

on ist (...) eine "Gleichung", bei der auf der linken Seite der durch eine Benennung ausgedrückte Begriff, das Definiendum, und auf der rechten Seite die Inhaltsbeschreibung des Begriffs, das Definiens, steht" (Arntz, et al., 2009: 59f.).

Im Rahmen dieser Arbeit wird auf den Wunsch der AVL hin darauf geachtet, die Formulierung der Definitionen so einfach wie möglich zu halten. Beispiel: "Glühlampe: ein materieller lichtaussendender Gegenstand (…), bei dem feste Stoffe durch Stromwärme so hoch erhitzt werden, dass sie Licht aussenden (…)" (Arntz et al., 2009:62).

6.2 Explikation

Schmitz (2003) beschreibt Explikation folgendermaßen: "Aussage, die einen Begriff beschreibt und ihn verständlicher macht, ihn allerdings nicht unbedingt von anderen Begriffen abgrenzt".

6.3 Kollokation

Schmitz (2003) definiert Kollokationen wie folgt: "Wiederkehrende im Zusammenhang stehende Wortkombination, deren Komponenten in einer Äußerung oder einer Reihe von Äußerungen zusammen auftreten, auch wenn diese nicht unbedingt in unmittelbarer Nähe zueinander stehen."

"Die Kollokation ist ein sprachliches Phänomen, das in der syntagmatischen Untersuchung der lexikalischen Ebene eine relevante Rolle spielt und das mit der typischen, konventionellen, rekurrenten Art der Kombination von Wortschatzelementen zu tun hat. Es handelt sich um das Problem präferierter Verbindbarkeit von Lexemen, um die Tatsache, dass einige Wortverbindungen ohne syntaktische oder semantische Regeln zu verletzen, nicht usuell sind, aber auch um die Tatsache, dass die Wahl eines Lexems die Wahl eines Partnerlexems festlegt (Rothkegel 1994:499f.). Als Beispiel für Kollokationen werden u.a. die folgenden gefunden: eingefleischter Junggeselle, den Tisch decken, starker Raucher, Kaffee trinken. Diese Beispiele haben gemeinsam, dass in ihnen die Verben oder Adjektive (also die Kollokatoren) keine oder wenige Synonyme aufweise und dass, auch wenn rein theoretisch das gleiche mit andren Adjektiven oder Verben ausgedrückt werden könnte (*Kaffee nehmen, *eingefleischter Raucher), dies aus Gründen der Norm im Sinne Coserius, also der Konvention nicht gängig ist" (Cedillo 2004, 31f.).

6.4 Kontext

"Im Sinne der Terminologiearbeit: Sprachliche oder außersprachliche Umgebung, in der eine Benennung oder eine Fachwendung auftreten kann" (DIN 2342-1, 1992)" (Soukup- Unterweger 2002). Schmitz (2003) charakterisiert Kontext als "Text oder Teil eines Textes, in dem eine Benennung erscheint". Arntz et al. (2009) definieren Kontext folgendermaßen: "Die Angabe des Kontextes sollte das Fachwort in seiner typischen Anwendung darstellen; damit kann der Kontext zugleich zum Verständnis der Bedeutung des Fachwortes beitragen."

Warburton (2008) definiert Kontext kurz und bündig wie folgt: "Text or part of a text in which a term occurs." Die Bedeutung von Kontexten im Rahmen der TE untersuchen Zielinksi und Safar (2005) im Zuge einer Onlineumfrage zum Einsatz von Terminologieextraktions- und Terminologieverwaltungstools: "Der Grund dafür kann darin gesehen werden, dass der Kontext für gewöhnlich eine sehr wichtige Rolle bei der Bestimmung der terminologischen Relevanz einer lexikalischen Einheit und damit für die Termerkennung spielt."

7. ProTerm

ProTerm ist das Werkzeug für der Firma ProCom Strasser. Im Vorfeld der Masterarbeit wurde ProTerm für "Dokumentation, Recherche, Pflege von Thesaurus, Verwaltung semantischer Netze, Navigation in Dokumenten, Beschlagworten von Dokumenten und Kommunikation intern und extern" (vgl. DocuMatrix 2007) verwendet. Folgende Formate können in ProTerm eingelesen und bearbeitet werden:

- .pdf
- .txt
- .html
- .xml
- alle Microsoft-Office-Formate.

Folgende Zeichensätze können in ProTerm eingelesen und bearbeitet werden:

- ISO 8859-1 bis 16
- VISC II
- ASC II
- UTF 8.

7.1 Vorbereitende Maßnahmen

Terminologie zu erkennen und sie als solche zu qualifizieren war eine der größten Herausforderungen im Rahmen dieser Masterarbeit. Die Autorin musste sich daher vor Beginn der TE mit den zur Verfügung gestellten Dokumentationen und den darin behandelten Fachgebieten vertraut machen, um TK ausfindig zu machen. Um so rasch wie möglich die TK identifizieren zu können, wurden vor Beginn der ersten TE die bestehenden Einträge der AVL-Terminologie-datenbank und der Text der Dokumentationsgruppe CAMEO gewissenhaft studiert. Somit konnte sich die Autorin in das Fachgebiet einarbeiten und sich ein Bild von dem geforderten Fachlichkeitsgrad der TK machen.

Ein Telefongespräch mit Elisabeth Stossier, die seit mehr als 20 Jahren für AVL als Übersetzerin tätig ist, hat dazu beigetragen, einen Einblick zu gewinnen, wie im Zuge der Übersetzungen neue Termini identifiziert, die zielsprachlichen Äquivalente

eruiert und Terminologie verwaltet werden. Frau Stossier arbeitet mit dem Übersetzungstool Transit (Firma STAR) und dem Textverarbeitungstool FrameMaker. Ihre zweisprachige Terminologiesammlung verwaltet sie in Microsoft Office Word- Listen, die sie bei Bedarf mit dem Suchtool TextPad nach bereits vorhandenen Übersetzungen durchsucht. Ihr großer Erfahrungsschatz im Umgang mit AVL-Dokumentationen trägt dazu bei in kurzer Zeit bislang neue Fachtermini zu erkennen.

7.2 Prozess Terminologieextraktion mit ProTerm

In dieser Masterarbeit werden die für die TE relevanten Arbeitsschritte in ProTerm detailliert beschrieben (siehe Abb. 1). Es soll gezeigt werden, wie mit ProTerm Terminologie und dazugehörige Datenelemente für die Datenkategorien extrahiert werden können. Das Terminologieextraktionsverfahren mit ProTerm kann in drei Phasen unterteilt werden:

- 1. Vorbereiten der Texte
- 2. Arbeit in ProTerm und Zuordnen der zielsprachlichen Äquivalente
- Export aus ProTerm

Zu Beginn werden die Texte vorbereitet (siehe Kapitel 7.2.1). Anschließend erfolgt die Arbeit mit ProTerm und somit die Phase der TE (siehe Kapitel 7.2.2 bis Kapitel 7.5). Den Abschluss der TE bildet der Export aus ProTerm (siehe Kapitel 7.7). Das Extraktionsverfahren, das bei ProTerm zur Anwendung kommt, wird den hybriden Extraktionsverfahren zugeordnet. Es ist eine Kombination der manuellen Extraktion (siehe Kapitel 4.3.1) und des statistischen Extraktionsverfahrens (siehe Kapitel 4.3.4).

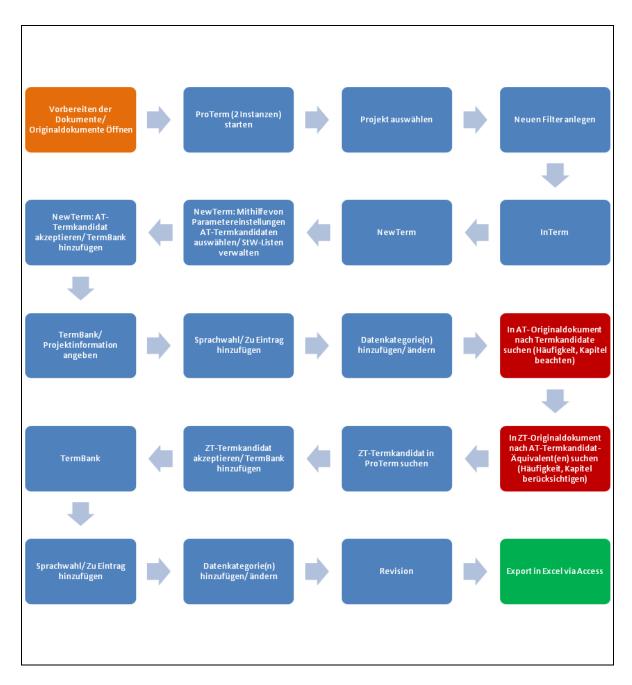


Abb. 1: Prozess Terminologieextraktion mit ProTerm

7.2.1 Vorbereiten der Texte

Die Texte wurden von AVL im pdf-Format zur Verfügung gestellt. Tabelle 1 zeigt, dass AVL Texte aus vier Dokumentationsgruppen bereit gestellt hat. Es wurde für jede Dokumentationsgruppe ein Ordner auf dem Desktop erstellt. Dieser Ordner enthält Unterordner für jede Sprache. Im Falle dieser Masterarbeit waren dies Ordner für Deutsch (de) und Englisch (en). In diese Ordner wurden danach die entsprechenden Dokumente nach Dokumentationsgruppen abgelegt. Vor dem Starten von ProTerm wurden die deutschen und englischen Texte geöffnet, um in der TE-Phase das Zuordnen von ausgangssprachlichen und zielsprachlichen Äquivalenten zu erleichtern (siehe Abb. 1).

7.2.2 ProTerm starten

Beim Start von ProTerm wird der Nutzer aufgefordert sich anzumelden (siehe Abb. 2). Es besteht die Möglichkeit beliebig viele Instanzen von ProTerm zu öffnen. Im Zuge dieser Masterarbeit wurde mit drei Instanzen (eine für die deutschen, eine für die englischen Texte und eine für beide Sprachen) gearbeitet (siehe Kapitel 7.2.5).

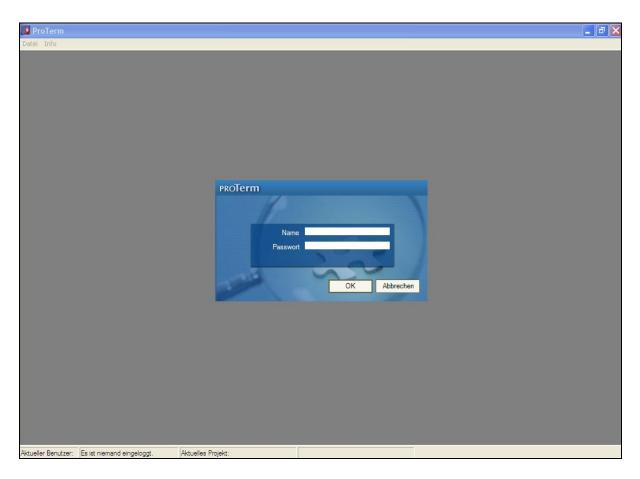


Abb. 2: ProTerm: Log-in

7.2.3 Projekt und Filter anlegen/auswählen

Nach dem ersten Einloggen in ProTerm wird ein Projekt angelegt, indem ein Projektname eingetragen wird (siehe Abb. 3). Ein Projekt legt fest, WAS eingelesen werden soll, also das "Thema". Ein Filter legt fest, WIE eingelesen werden soll, also wie auf Daten, durch Einstellen von Parametern, zugegriffen werden soll. Innerhalb eines Projekts können mehrere Filter festgelegt werden.

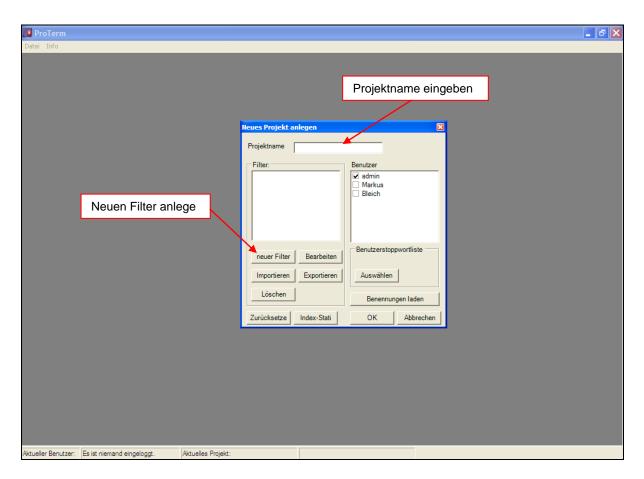


Abb. 3: ProTerm: Neues Projekt anlegen

Nachdem ein neues Projekt angelegt ist, wird ein Filter erstellt. In diesem Filter wird festgelegt, wo sich die einzulesenden Dokumente befinden, welche Formate sie haben (siehe Kapitel 7.2.1) und welche StW-Listen beim Einlesen berücksichtigt werden sollen. StW-Listen dienen dazu festzulegen, welche Termini während dem Einlesen ausgeschlossen werden und somit im Einleseergebnis nicht aufscheinen. Von den Entwicklern wurden StW-Listen mit allgemeinsprachlichen deutschen und englischen Termini zur Verfügung gestellt. Diese werden auch von den gängigen Internet-Suchmaschinen verwendet. Vor dem ersten Einlesen wurden StW-Listen erstellt, die die bereits existierenden Einträge in der AVL-Terminologiedatenbank beinhalten, um zu gewährleisten, dass diese Einträge nicht erneut angezeigt werden. Beim Erstellen des Filters werden außerdem das einzulesende Format, der Zeichensatz und eine Liste der Sonderzeichen festgelegt. Es besteht auch die Möglichkeit, beim Erstellen des Filters festzulegen, dass Zahlen, kleingeschriebene Wörter oder HTML-Tags während des Einlesevorgangs entfernt werden, sowie die minimale beziehungsweise maximale Länge der Zeichen vorzugeben (siehe Abb. 4).

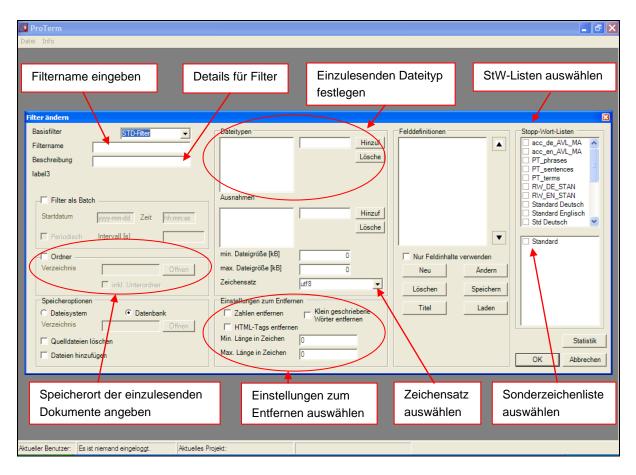


Abb. 4: ProTerm: Neuen Filter anlegen

Für jede AVL-Dokumentationsgruppe wurde in ProTerm ein Projekt angelegt, dem jeweils drei Filter zugeteilt wurden: ein Filter für die deutschen Texte, ein Filter für die Texte in englischer Sprache und ein dritter Filter für alle Texte einer AVL-Dokumentationsgruppe (siehe Abb. 5 und Abb. 6).

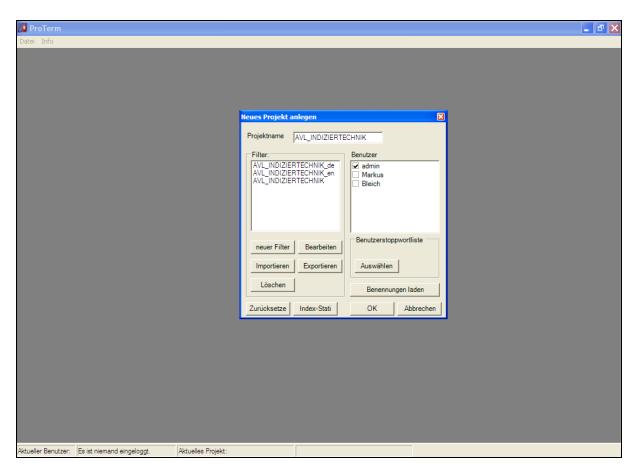


Abb. 5: ProTerm: Projekt AVL-Indiziertechnik und Filter

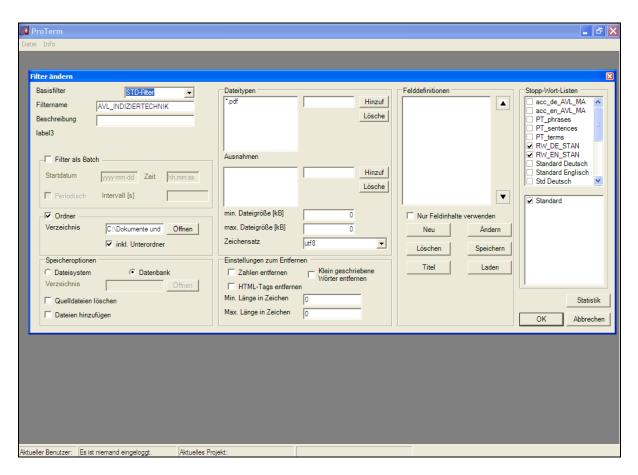


Abb. 6: ProTerm: Filter AVL_Indiziertechnik

7.2.4 InTerm Einlesen der Dokumente

Das Einlesen der Dokumente erfolgt über die Aktivierung des Filters. Nachdem Projekt und Filter angelegt wurden, erscheint ein Fenster, das dazu auffordert den zuletzt bearbeiteten Filter zu aktivieren (siehe Abb. 7). Filter können auch in Administration → Projekteditor (siehe Abb. 52) bearbeitet werden. Nachdem der Filter angelegt wurde, ist er durch Klicken auf das Icon InTerm am linken Bildrand zu starten. In diesem Bereich besteht die Möglichkeit zwischen den Filtern eines Projektes zu wechseln und die jeweiligen Einleseparameter zu überprüfen und gegebenenfalls zu adaptieren (siehe Abb. 8). ProTerm kann pdf-, txt-, html-, xml- sowie alle Microsoft-Office-Formate einlesen. Im Zuge dieser Masterarbeit wurde ausschließlich mit .pdf-Formaten gearbeitet. Nachdem der Einlesevorgang beendet ist, erstellt ProTerm eine Übersicht über das Ergebnis des Einlesevorgangs. Es werden Anzahl der eingelesenen Dateien, der gefundenen Benennungen, der neuen Benennungen und der bestehenden Benennungen sowie die Laufzeit des Einlesevorgangs angezeigt (siehe Abb. 9). Nach Rücksprache mit Experten stellte sich heraus, dass ProTerm im Vergleich zu anderen TET weniger Zeit zum Einlesen von Dokumenten benötigt.

Eine Übersicht über die Dauer des Einlesevorgangs und über die bestehenden und gefundenen Benennungen ist Abb. 9 und Tab. 2 zu entnehmen.

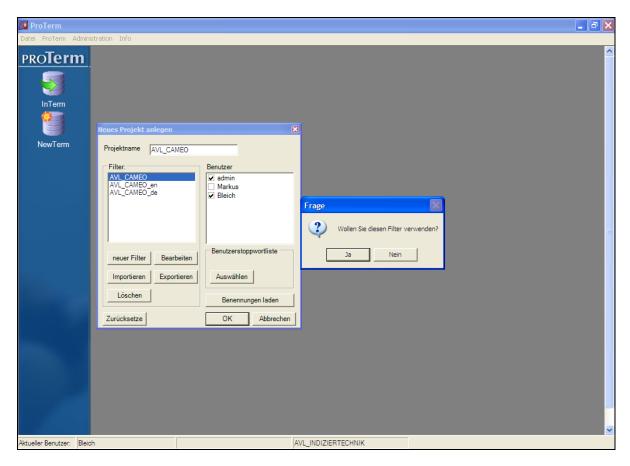
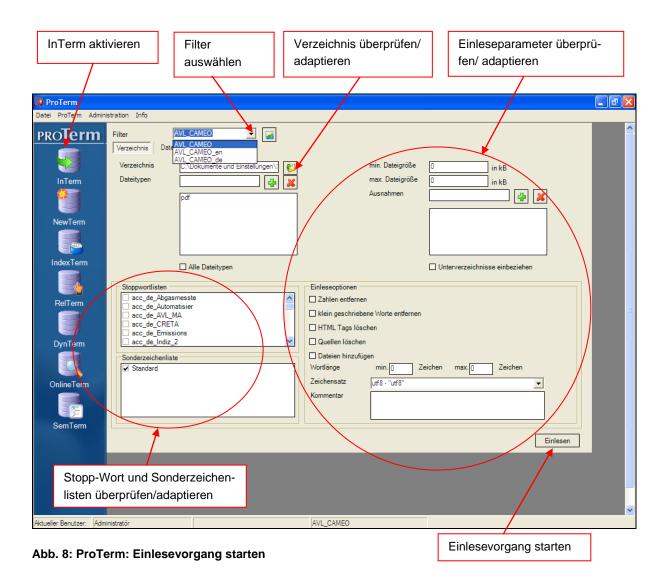


Abb. 7: Filter aktivieren



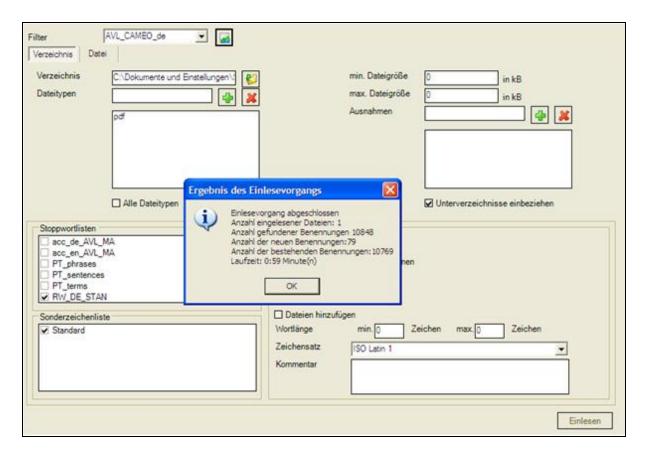


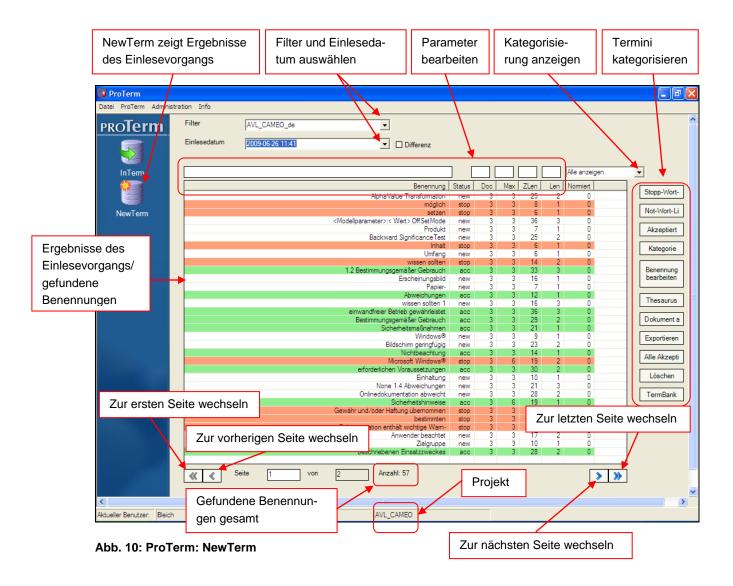
Abb. 9: ProTerm: Ergebnis des Einlesevorgangs

Tab. 2: Übersicht Einlesedauer

Projekt/						Dauer des
Dokumentations-				Anzahl der	Seitenanzahl	Einlesevorgangs
gruppe	Sprache(n)	Datum	Uhrzeit	Dokumente	(pdf-Format)	(in Minuten)
EMCON	de	03.08.09	13:35	4	448	05:57
EMCON	en	03.08.09	13:45	4	448	07:24
EMCON	de, en	03.08.09	13:47	8	896	04:08
SANTORIN	de	25.08.09	20:50	3	390	00:59
SANTORIN	en	25.08.09	20:51	3	384	00:46
SANTORIN	de, en	25.08.09	20:52	6	774	01:24
INDIZIERTECHNIK	de	27.08.09	01:09	1	212	00:34
INDIZIERTECHNIK	en	27.08.09	01:12	1	208	00:32
INDIZIERTECHNIK	de, en	27.08.09	01:13	2	420	01:01

7.2.5 NewTerm Terminologieextraktion

Nach den vorbereitenden Schritten erfolgt die TE. Die Projekte und Filter wurden angelegt und eingelesen. Durch Klicken auf das Icon NewTerm am linken Bildschirmrand kann das zu bearbeitende Projekt und der dazugehörige Filter anhand des Einlesedatums ausgewählt werden. ProTerm erstellt eine Liste mit den Benennungen. Diese Benennungen sind Tokens, vor und/oder nach deren Auftreten sich im Text ein Stopp-Wort und/oder ein Sonderzeichen befindet (siehe Abb. 10). Es werden nicht nur Einwortbenennungen dargestellt, sondern auch Mehrwortbenennungen. Als Ausgangssprache für die TE wurde Deutsch gewählt, da es die Ausgangssprache der Originaldokumente sowie die Muttersprache der Autorin ist, was die Identifikation von TK erleichterte. Durch Doppelklicken auf eine Benennung oder Markierung einer Benennung und Klicken auf das Icon Dokument anzeigen am rechten Bildschirmrand kann in die Dokumentansicht (siehe Abb. 11) und anschließend in das Originaldokument gewechselt werden. In der Dokumentansicht kann mittels der Suchfunktion Strg+F nach einer beliebigen Benennung gesucht werden. Diese Funktion und das Wechseln in die Originaltextansicht ermöglichen es dem Terminologen sich während des Extraktionsprozesses ein Bild von der Umgebung des TK zu machen. Die in NewTerm ausgewählte Benennung ist farblich hervorgehoben (siehe Abb. 11).



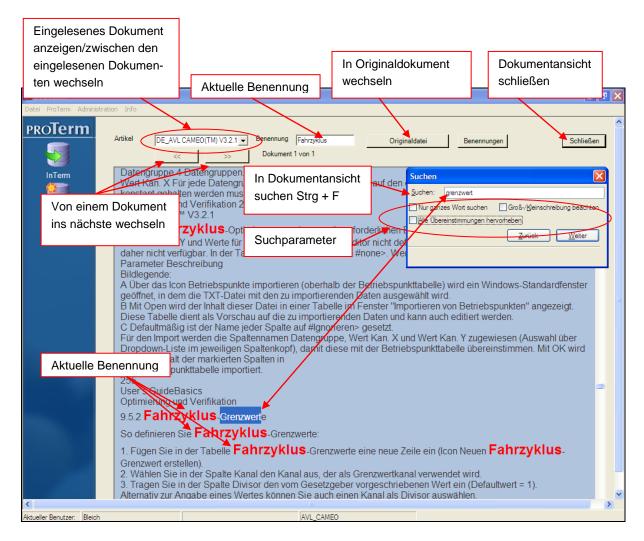


Abb. 11: ProTerm: Dokumentansicht, Suche in Dokumentansicht

7.2.5.1 Parameter

Die Parameter dienen dazu das Einleseergebnis zu filtern. In Abb. 12 sind die zur Verfügung stehenden Parameter dargestellt und ihre Funktion wird im Anschluss detailliert beschrieben.

] [Alle anzeig	gen	_
Benennung	Status	Doc	Max	ZLen	Len	Nomiert		Г
System	acc	- 1	- 1	6	- 1	0		L
Beschreibung	stop	2	3	12	- 1	0		_
Anzeige	new	2	8	7	1	0		
AVL List GmbH	new	1	1	13	3	0		

Abb. 12: ProTerm: NewTerm- Parameter

a. Benennung

Im Feld *Benennung* kann eine Benennung gesucht werden. Mithilfe des Trunkierungssymbols * (vor und/oder nach der Benennung) werden Benennungen gefunden, die aufgrund von ihrer Position im Text⁷ nicht als Einzelwort aufgelistet werden (siehe Abb. 13). Das Auffinden von Mehrwort-Kombinationen erfolgt durch das gemeinsame Verwenden des Trunkierungssymbols * (vor und/oder nach der Benennung), der Leertaste und einem der beteiligten Termini (siehe Abb. 14). Durch Klicken auf das Icon *Benennung* werden die gefundenen Tokens alphabetisch (initialalphabetisch oder finalalphabetisch) sortiert. Die Trunkierungsfunktion kann dazu genutzt werden einheitliche Terminologie zu extrahieren (siehe Abb. 15).

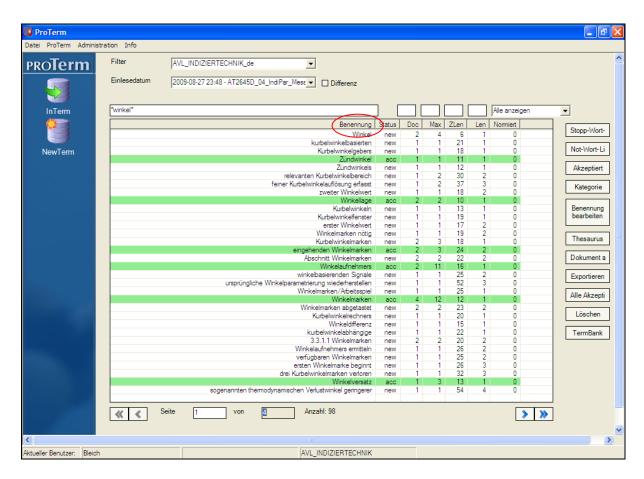


Abb. 13: ProTerm: NewTerm- Benennung - Trunkierung *winkel*

⁷ Steht eine gesuchte Benennung nicht vor oder nach einem Stopp-Wort beziehungsweise Sonderzeichen, wird sie hier nicht als Einzelwort angezeigt.

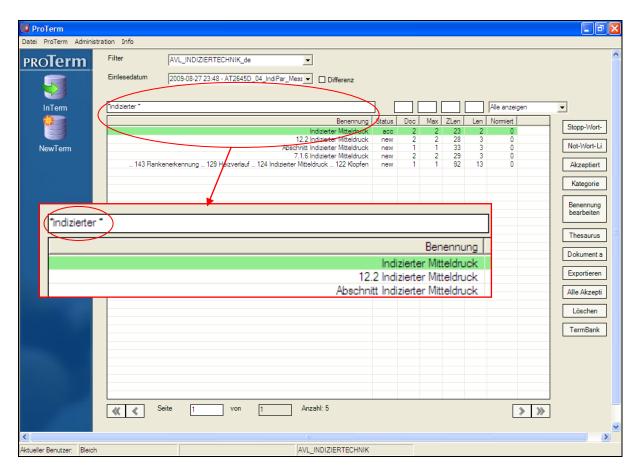


Abb. 14: ProTerm: NewTerm- Benennung – Trunkierung *indizierter_*

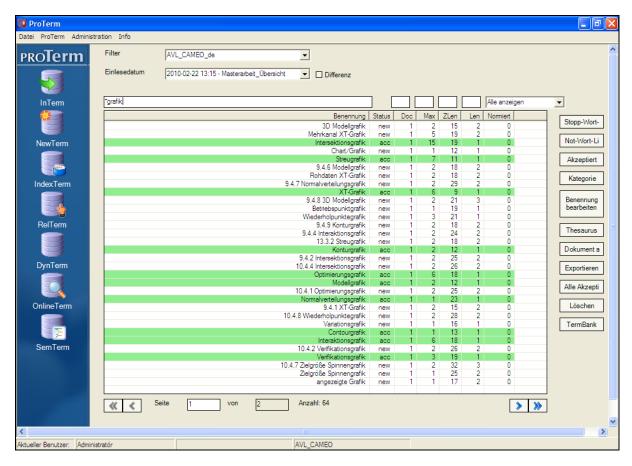


Abb. 15: ProTerm: NewTerm- Benennung - Trunkierung *grafik

b. Status

Durch Klicken auf den Spaltenkopf *Status* kann nach den Kategorisierungen sortiert werden (siehe Abb. 16). Noch nicht kategorisierte Benennungen haben den Status *new*. Als Stopp-Wort qualifizierte Benennungen werden als *stop* und akzeptierte Benennungen als *acc* angezeigt.

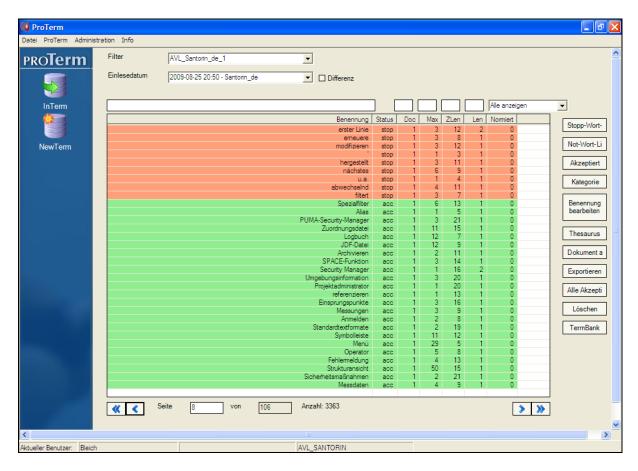


Abb. 16: ProTerm: NewTerm - Gefiltert nach Status

c. Doc

In diesem Feld kann nach der Anzahl der Dokumente gefiltert werden. Durch Klicken auf den Spaltenkopf *Doc* können die Ergebnisse in aufsteigender oder absteigender Reihenfolge angezeigt werden (siehe Abb. 17). Mithilfe der Verhältniszeichen *Größer als* > und *Kleiner als* < ist es möglich das Ergebnis zu verfeinern (siehe Abb. 18).

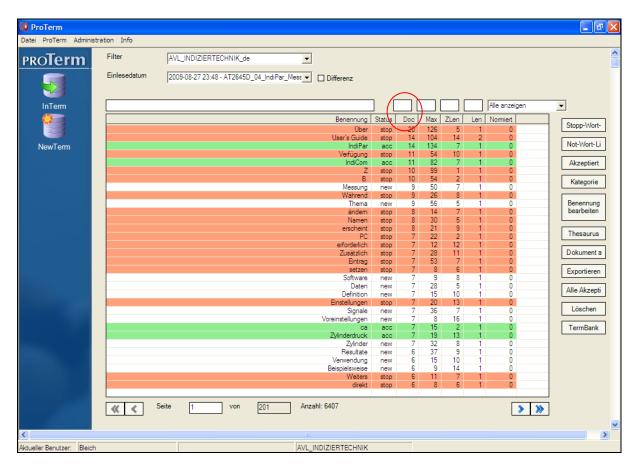


Abb. 17: ProTerm: NewTerm - Gefiltert nach Häufigkeit der Dokumente (absteigend)

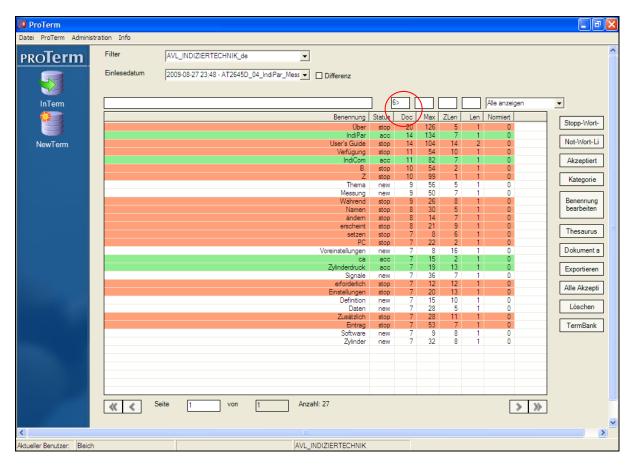


Abb. 18: ProTerm: NewTerm – Gefiltert nach Häufigkeit der Dokumente in mehr als sechs eingelesenen Dokumenten

d. Max

In diesem Feld kann nach der absoluten Häufigkeit der Benennungen in den eingelesenen Dokumenten gefiltert werden. Durch Klicken auf den Spaltenkopf *Max* können die Ergebnisse in aufsteigender oder absteigender Reihenfolge angezeigt werden (siehe Abb. 19). Mithilfe der Verhältniszahlen *Größer als* > und *Kleiner als* < ist es möglich das Ergebnis zu verfeinern (siehe Abb. 20).

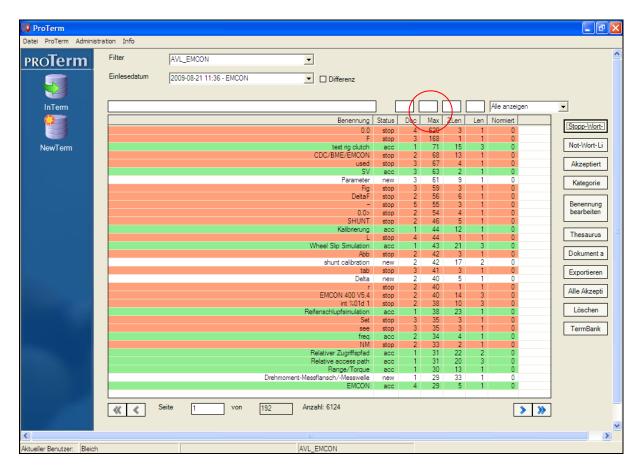


Abb. 19: ProTerm: NewTerm - Max - Gefiltert nach Häufigkeit (absteigend)

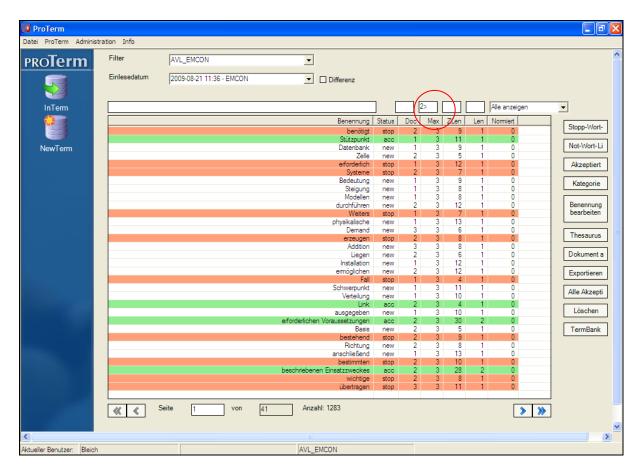


Abb. 20: ProTerm: NewTerm - Gefiltert nach Benennungen, die öfter als zweimal auftreten

e. ZLen

In diesem Feld kann nach der Anzahl der Zeichen einer Benennung gefiltert werden. Durch Klicken auf den Spaltenkopf *ZLen* können die Ergebnisse in aufsteigender oder absteigender Reihenfolge angezeigt werden. Mithilfe der Verhältniszeichen *Größer als* > und *Kleiner als* < ist es möglich das Ergebnis zu verfeinern (siehe Abb. 21).

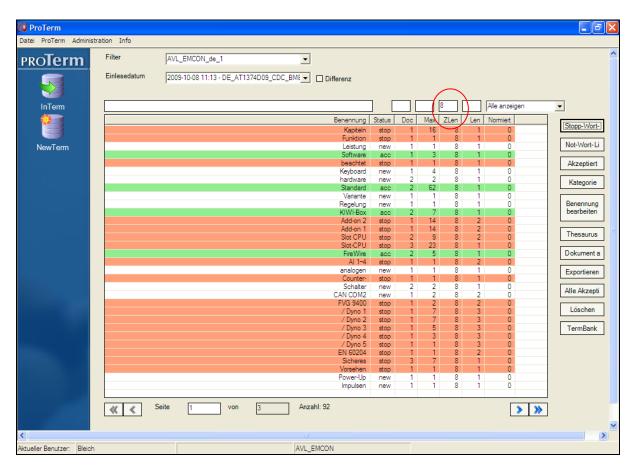


Abb. 21: ProTerm: NewTerm – Benennungen mit acht Zeichen werden angezeigt

f. Len

In diesem Feld kann nach Mehrwortsequenzen gefiltert werden (siehe Abb. 22). Durch Klicken auf den Spaltenkopf *Len* können die Ergebnisse in aufsteigender oder absteigender Reihenfolge angezeigt werden. Mithilfe der Verhältniszeichen *Größer als* > und *Kleiner als* < ist es möglich das Ergebnis zu verfeinern.

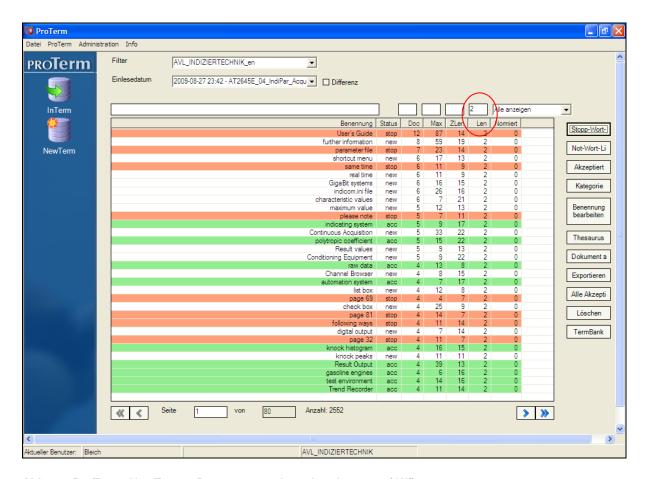


Abb. 22: ProTerm: NewTerm - Benennungen bestehend aus zwei Wörtern

g. Normiert

Durch das Betätigen dieses Schalterknopfes kann nach den zugewiesenen Normierungen gefiltert werden. Im Zuge dieser Masterarbeit wurde dieses Feld nicht verwendet und soll daher nur der Vollständigkeit halber erwähnt werden.

h. Auswahlmenü

Das Auswahlmenü ermöglicht die Filterung nach den zugewiesenen Kategorisierungen (Alle Anzeigen, Neu, Akzeptiert, Not-Wort-Liste, Stopp-Wort-Liste, Normierte, Thesaurus) (siehe **Fehler! Verweisquelle konnte nicht gefunden werden.**). Nach Auswahl einer der Kategorisierungen wird ausschließlich diese angezeigt. Im Zuge dieser Arbeit wurde immer mit Alle anzeigen gearbeitet, da so dank der Farbkodierungen (siehe **Fehler! Verweisquelle konnte nicht gefunden werden.**) ersichtlich war, welche Benennung im Vorfeld welcher Kategorie zugewiesen wurde.

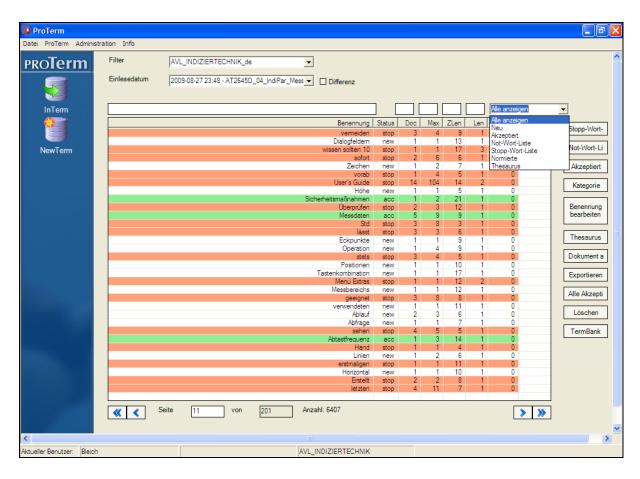


Abb. 23: ProTerm: NewTerm – Auswahlmenü, Farbkodierung

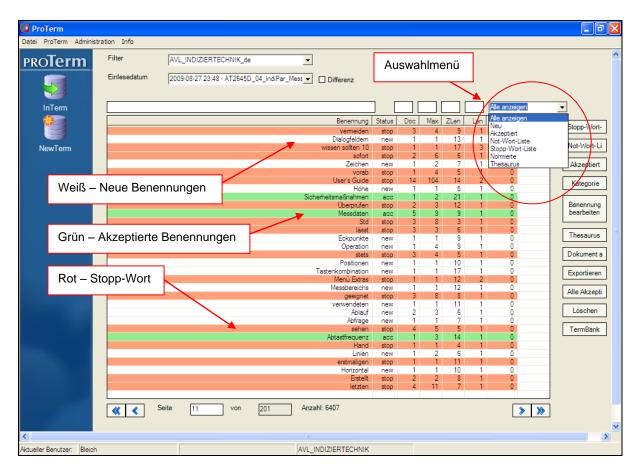


Abb. 23: ProTerm: NewTerm - Auswahlmenü, Farbkodierung

i. Farbkodierung

Neue Benennungen sind in weißer Farbe hinterlegt, akzeptierte Termini in grüner und Stopp-Wörter in roter Farbe (siehe **Fehler! Verweisquelle konnte nicht gefunden werden.**). Werden die Benennungen nach dem Status sortiert oder wird im Auswahlmenü eine Kategorisierung ausgewählt, so werden die Benennungen in den jeweiligen Farben hinterlegt angezeigt.

7.2.5.2 Kombinieren der Parameter

Die Parameter *Benennung*, *Status*, *Doc*, *Max*, *ZLen*, *Len* können individuell miteinander kombiniert und parallel angewendet werden. Dies ermöglicht es dem Terminologen, das Einleseergebnis nach Belieben zu filtern und Einschränkungen für das Anzeigen des Einleseergebnisses aufzustellen.

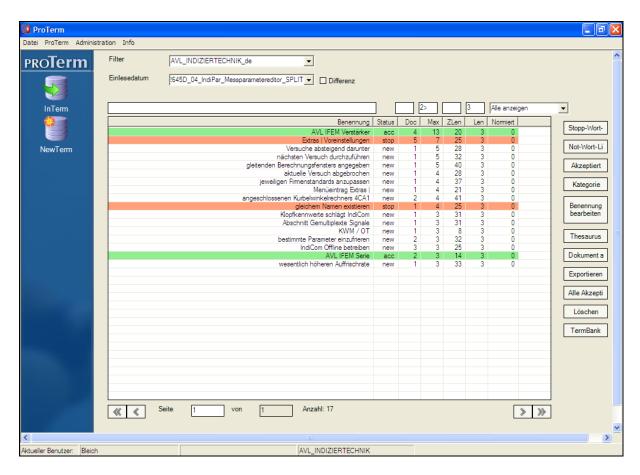


Abb. 24: ProTerm: NewTerm – Zeige Drei-Wort-Benennungen, die öfter als zweimal auftreten nach ihrer absoluten Häufigkeit sortiert

7.3 Hinzufügen von Termkandidaten

Hier wird erläutert, wie ein Termkandidat⁸ ausgewählt wird. Dies erfolgt durch das Akzeptieren einer Benennung in NewTerm. Nachdem der NewTerm-Bereich aktiviert ist und ein Termkandidat identifiziert wurde (zur Identifikation von TK siehe Kapitel 7.4), wird dieser markiert. Durch Betätigen der Funktionstaste *F5* oder durch Klicken auf die Schaltfläche Akzeptiert wird der Termkandidat lediglich akzeptiert und ist der TermBank im Anschluss durch Klicken auf das TermBank-Icon hinzuzufügen. Durch Markieren des zukünftigen Termkandidaten und Klicken auf die Schaltfläche Term-Bank wird dieser akzeptiert und direkt der TermBank hinzugefügt. Es öffnet sich das SelectSentence-Fenster (siehe Abb. 25), in dem alle Textpassagen, in denen der Termkandidat vorkommt, aufgelistet werden. Die Darstellungsform ist auf Sätze voreingestellt, d. h. es werden ganze Sätze angezeigt. Es besteht die Möglichkeit in die Phrasenansicht, in der nur Textteile angezeigt werden, zu wechseln (siehe Abb. 26). Der ausgewählte Termkandidat wird im Feld Token angezeigt. Mithilfe des Feldes Suchen wird die Suche innerhalb der Textpassagen, in denen der TK vorkommt, beschleunigt. Durch Auswählen von Sätzen oder Phrasen kann die Darstellungsform der Textpassagen geändert werden. Im SelectSentence-Fenster kann eine Datenkategorie hinzugefügt werden. Die Datenkategorie Kontext ist vorausgewählt, da sie am häufigsten zur Anwendung kam. Es besteht die Möglichkeit, durch Klicken auf eine andere Datenkategorie, diese auszuwählen. In der TermBank können weitere Datenkategorien ergänzt werden. Es ist notwendig eine Datenkategorie auszuwählen, auch wenn diese im Anschluss wieder gelöscht wird, da nur so die Quelle des Termkandidaten übernommen werden kann. Durch Klicken auf eine der Textpassagen (siehe Abb. 25 und Abb. 26) wird diese vollständig im unteren Feld angezeigt. Durch Doppelklicken auf eine Textpassage oder Klicken auf das Icon Text kann in das eingelesene Dokument, also in die Dokumentansicht, und anschließend in das Originaldokument gewechselt werden. Im eingelesenen Dokument kann mittels der Suchfunktion Strg+F nach einer beliebigen Benennung gesucht werden. Die in New-Term ausgewählte Benennung ist farblich hervorgehoben (vgl. Abb. 11). Wird allerdings keine Datenkategorie ausgewählt, so kann manuell eine Quelle eingegeben werden. Sollte keine Quelle eingetragen werden, wird der TK trotzdem der Term-Bank hinzugefügt. Unabhängig davon, ob eine Datenkategorie ausgewählt oder eine Quelle angegeben wurde, erscheint das *AddNewTerm*-Fenster (siehe Abb. 27).

Hier wird die Sprache ausgewählt und angegeben, ob in der *TermBank* ein neuer Eintrag anzulegen oder der TK einem bestehenden Eintrag hinzuzufügen ist. Das

_

⁸ Sobald eine Benennung der *TermBank* hinzugefügt wird, wird sie als *Termkandidat* bezeichnet.

Auffinden eines bestehenden Eintrags ist mithilfe der Trunkierungsfunktion möglich. In der TermBank wird nun der neu angelegte oder der zu einem bestehenden Eintrag hinzugefügte TK angezeigt. Wurde, wie bereits erwähnt, keine Quelle angegeben und ist somit eine "leere" Datenkategorie vorhanden, so ist dieses Feld in blassroter Farbe hinterlegt (siehe Abb. 28). Die farblich hinterlegten, leeren Felder erleichtern das Auffinden "leerer" Datenkategorien, die vor dem Export zu löschen sind, um einen Mehrfachexport zu vermeiden. In der TermBank können nach Belieben weitere Datenkategorien hinzugefügt werden. Durch Klicken auf das Icon Hinzufügen öffnet sich erneut ein SelectSentence-Fenster. Als Voraussetzung für einen erfolgreichen Export der Termkandidaten aus der ProTerm-TermBank ist eine Projekt-Information anzugeben und darauf zu achten, dass, wenn ein TK in nur einer Sprache verfügbar ist, weil er beispielsweise nicht übersetzt wurde, in der anderen Sprache ein "leerer" Eintrag anzulegen und eventuell mit einer Anmerkung zu versehen ist. Das Hinzufügen eines "leeren" Eintrages erfolgt durch Markieren des bestehenden Eintrages und anschließendem Klicken auf Hinzu. Das nun erscheinende Feld wird nicht beschriftet. Nachdem mit OK bestätigt wurde, erscheint das AddNewTerm-Fenster, in dem die nicht vorhandene Sprache ausgewählt wird. Dieses Fenster bietet zudem die Möglichkeit zu kontrollieren, ob dem korrekten Eintrag ein "leerer" Eintrag hinzugefügt wird. Sollte dies nicht der Fall sein, kann hier ein anderer Eintrag ausgewählt werden. Das Vorhandensein von Einträgen in beiden Sprachen ist für den Export der TK aus ProTerm notwendig (siehe Kapitel 7.7).

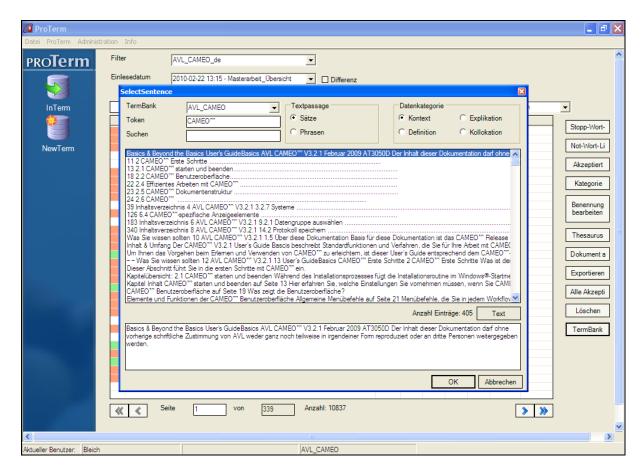


Abb. 25: ProTerm: SelectSentence- Fenster – Textpassage Sätze

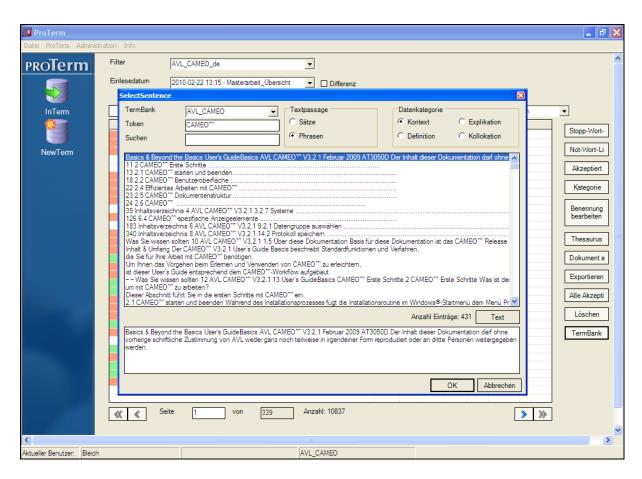


Abb. 26: ProTerm: SelectSentence- Fenster – Textpassage Phrasen

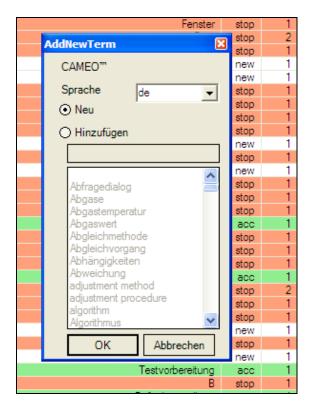


Abb. 27: ProTerm: AddNewTerm-Fenster

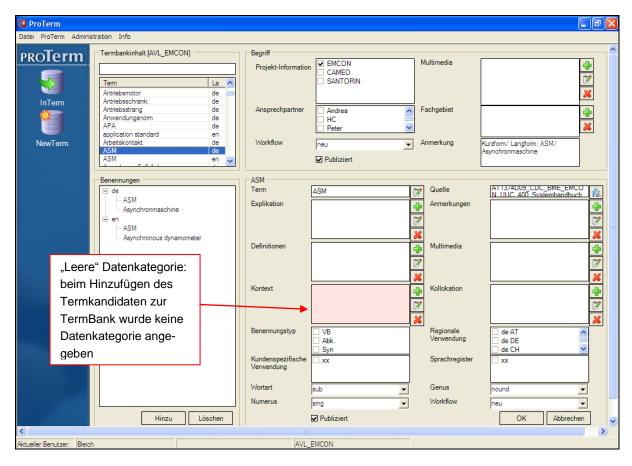


Abb. 28: ProTerm: Neuer Eintrag in TermBank

7.4 Identifizieren von Termkandidaten

Zerfaß (2006) stellt in den Raum, dass "die Extraktion von Terminologie eine sehr subjektive Angelegenheit ist" und vom "Kenntnisstand der Materie" des Terminologen und der "Zielsetzung für die weitere Verwendung der Terminologie" Zerfaß (2006) beeinflusst ist. Eine der AVL-Vorgaben lautete, so viele Fachtermini wie möglich zu extrahieren. Diese Fachtermini sollten einen möglichst hohen Fachlichkeitsgrad aufweisen und so viele Fachgebiete wie möglich abdecken. Um so rasch wie möglich die gewünschten Fachtermini identifizieren zu können, wurden vor Beginn der ersten TE die bestehenden Einträge der AVL-Terminologiedatenbank und der Text der Dokumentationsgruppe CAMEO gewissenhaft studiert. Dies sollte dazu beitragen, dass während der TE ohne großen Zeit- und Rechercheaufwand und mithilfe der Parametereinstellungen Termkandidaten ausfindig gemacht werden könnten. Eine weitere Hilfe bei der Auswahl der TK war eine intensive Auseinandersetzung mit dem Inhalts-, dem Index- und, wenn vorhanden, dem Abkürzungsverzeichnis.

Diese Verzeichnisse ließen ebenfalls auf potentielle TK schließen und wurden zu Beginn jeder Dokumentationsgruppe untersucht. Anschließend wurde mithilfe der Parameter, wie in Kapitel 7.2.5.1 beschrieben, gearbeitet. Zuerst wurden, wie bereits erwähnt, die deutschen Texte bearbeitet und anschließend wurden die englischen Texte auf mögliche, im Deutschen nicht ersichtliche TK durchsucht. Das Auffinden der jeweiligen zielsprachlichen Äquivalente erfolgte durch Einsicht in die Originaldokumente. Dabei wurde berücksichtigt, in welchen Kapiteln und, je nach Dokumentation, auf welcher Seite der Termkandidat verwendet wurde. Diese Textstellen wurden dann in den entsprechenden Textpassagen der Zielsprache auf mögliche Termkandidaten durchsucht, die wiederum im Anschluss in ProTerm gesucht und dem ausgangssprachlichen TK hinzugefügt wurden (siehe Kapitel 7.3)⁹. Um die Suche noch rascher zu gestalten, wäre es von Vorteil, wenn bei der Weiterentwicklung von ProTerm das Auffinden der zielsprachlichen Kapitel automatisiert werden würde (Details dazu siehe Kapitel 0).

7.5 Revision

Bei der Revision wurde in der *TermBank* kontrolliert, ob für jeden Eintrag beide Sprachen vorhanden waren. Wie in Kapitel 7.3 erwähnt, ist es für den Export aus ProTerm wichtig, dass jeder Termkandidat ein Äquivalent in der anderen Sprache hat. Handelt es sich um eine Nullübersetzung, wurde ein "leerer Eintrag" in der anderen Sprache angelegt. Nur so konnte der Export erfolgreich durchgeführt werden. Bei der Revision wurde zudem darauf geachtet, dass keine farblich hinterlegten Datenkategorien vorhanden waren, da dies zu einem Mehrfachexport geführt hätte. Um die Projekte den Dokumentationsgruppen zuzuordnen, wurde das Feld *Projekt- Information* auf Begriffsebene der *TermBank* dazu genutzt, zwischen den Dokumentationsgruppen zu unterscheiden.

7.6 TermBank

Die ProTerm *TermBank* ist eine begriffsorientierte Terminologiedatenbank. Sie besteht aus einem Übersichtsfeld aller Einträge (1) und drei Ebenen: der Begriffsebene (2), der Sprachebene (3) und der Termebene (4) (siehe Abb. 29). Im *TermBank Manager* (*Administration* → *TermBank Manager*, siehe Abb. 52) können Kategorien hinzugefügt, geändert oder entfernt werden (siehe Abb. 30). In Tab. 3 wird ein Überblick über die im Zuge der TE mit ProTerm verwendeten Icons gegeben.

⁹ Bei den in dieser Arbeit behandelten Texten handelte es sich um technische Fachtexte, die systematisch erstellt wurden, was das Auffinden der jeweiligen zielsprachlichen Äquivalente mithilfe der Kapitel- beziehungsweise Seitenangabe erheblich erleichterte.

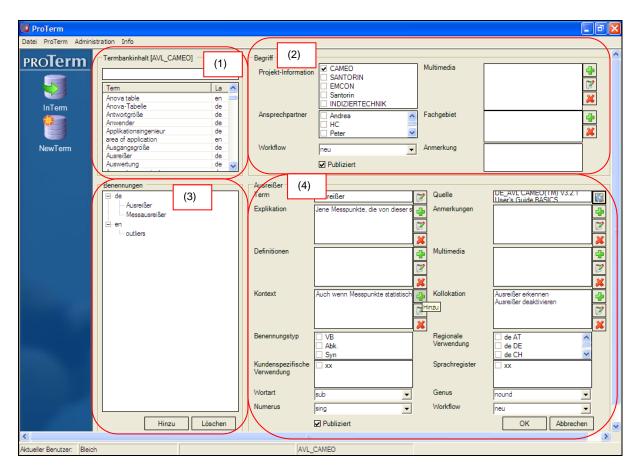


Abb. 29: ProTerm: TermBank

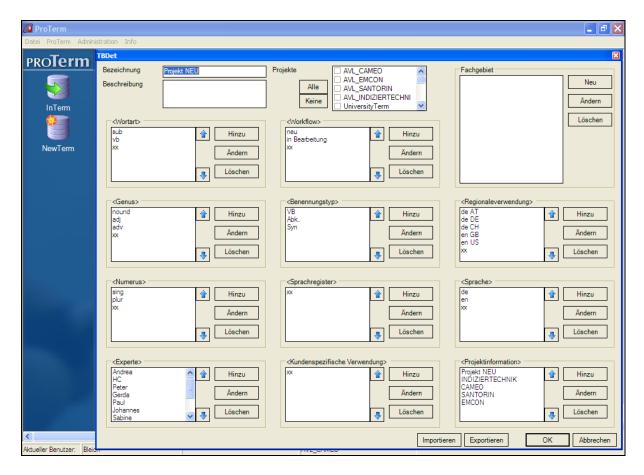


Abb. 30: ProTerm: TermBank Manager

Tab. 3: ProTerm: TermBank - Icons

Icon	Anwendung				
	Ändern				
4	Hinzufügen				
*	Löschen				
į.	Öffnen				

7.6.1 Termbankinhalt

Im Übersichtsfeld werden alle Einträge der *TermBank* in alphabetischer Reihenfolge aufgelistet. Im Suchfeld kann mithilfe der Trunkierungsfunktion mit * ein Eintrag gesucht werden (siehe Abb. 31).

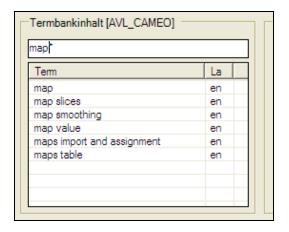


Abb. 31: ProTerm: TermBank - Termbankinhalt

7.6.2 Begriffsebene

Auf der Begriffsebene können *Projekt-Information*, *Ansprechpartner*, der *Workflow* Status oder das *Fachgebiet* angegeben werden. Zudem besteht die Möglichkeit eine Multimedia-Datei hochzuladen oder eine Anmerkung zu schreiben (siehe Abb. 32). Im Zuge dieser Masterarbeit wurden auf Begriffsebene ausschließlich die Felder *Projekt-Information* und *Anmerkung* verwendet. Das Feld *Projekt-Information* ist für einen reibungslosen Export aus ProTerm notwendig (siehe Kapitel 7.7). Je nach Projekt kann im *TermBank Manager* eine weitere Projekt-Information hinzugefügt werden. Anmerkungen, die auf der Begriffsebene gemacht werden, gelten für die Benennungen in allen Sprachen.



Abb. 32: ProTerm: TermBank - Begriffsebene

7.6.3 Sprachebene

Auf der Sprachebene können der jeweiligen Sprache beliebig viele Benennungen hinzugefügt oder vorhandene Einträge gelöscht werden. Im *TermBank Manager* können Sprachen hinzugefügt, durch Klicken auf die rechte Maustaste geändert oder gelöscht werden (siehe Abb. 33).

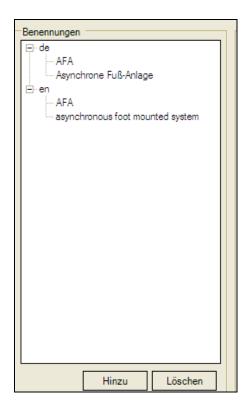


Abb. 33: ProTerm: TermBank - Sprachebene

7.6.4 Termebene

Auf der Termebene wird der aktuell ausgewählte Term angezeigt und er kann manuell nachbearbeitet werden. Das Bearbeiten einer Benennung ist eine nützliche Funktion, die zur Anwendung kommt, wenn Termini nicht in ihrer Grundform in die TermBank übernommen wurden (zum Beispiel bei Termini im Plural, flektierten Termini, o. Ä.). Die Quelle zu jedem Term wird von ProTerm automatisch übernommen, unter der Voraussetzung, dass während des Extraktionsprozesses eine Datenkategorie ausgewählt wurde, von der die Quelle dann für den betreffenden Term übernommen werden kann. Es besteht die Möglichkeit, aus den eingelesenen Texten die Datenelemente für die Datenkategorien (*Explikation*, *Definition*, *Kontext* und *Kollokation*) hinzuzufügen, zu bearbeiten oder zu löschen. Ein Anmerkungsfeld auf Termebene dient dazu etwaige, den Term betreffende Kommentare zu machen. Die Felder *Multimedia*, *Benennungstyp*, *Kundenspezifische Verwendung*, *Wortart*, *Numerus*, *Regionale Verwendung*, *Sprachregister*, *Genus*, *Workflow* und *Publiziert* wurden im Rahmen dieser Masterarbeit nicht verwendet und seien hiermit nur der Vollständigkeit halber erwähnt (siehe Abb. 34).

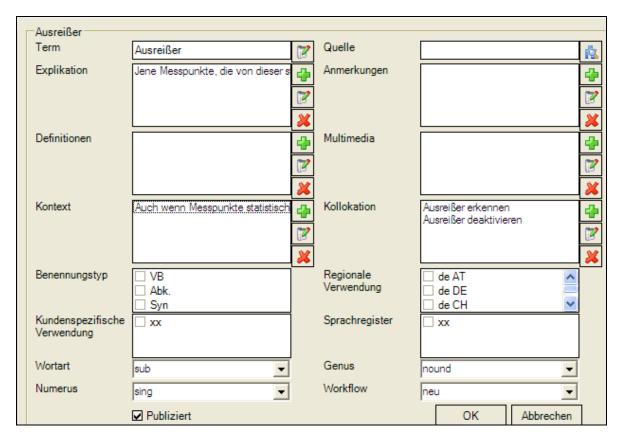


Abb. 34: ProTerm: TermBank - Termebene

7.7 Export aus ProTerm

Der Export der Extraktionsergebnisse aus ProTerm erfolgte über eine eigens von Amtsdirektor Pilles erstellte Access-Abfrage in eine Excel-Tabelle. Bei Projektstart wurde von AVL eine Muster-Excel-Tabelle zur Verfügung gestellt, die als Vorlage für den Export via Access dienen sollte. Diese Excel-Tabelle war nach AVL-Vorgabe für den Import der TK in die AVL-Termdatenbank generiert. Um die in ProTerm angelegten Projekte nun wieder den Dokumentationsgruppen zuzuordnen, wurde in Access die Projekt-Information angegeben (siehe Abb. 35 und Abb. 36). Auf Wunsch von AVL wurde hier auch das Feld *Sachgebiet* mit "not yet assigned" vorbelegt. Der Export erfolgte in eine Word-RTF-Datei (siehe Abb. 37). Die Exportergebnisse wurden in dieser Datei gesamt markiert, kopiert und anschließend in eine leere Excel-Tabelle eingefügt. Diese Excel-Tabelle wurde zum Abschluss benutzerfreundlich formatiert und an AVL übergeben (siehe Abb. 38).

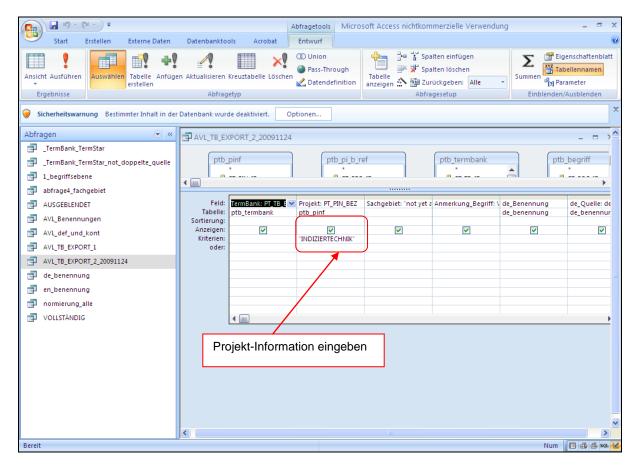


Abb. 35: Microsoft Access: Export der Extraktionsergebnisse

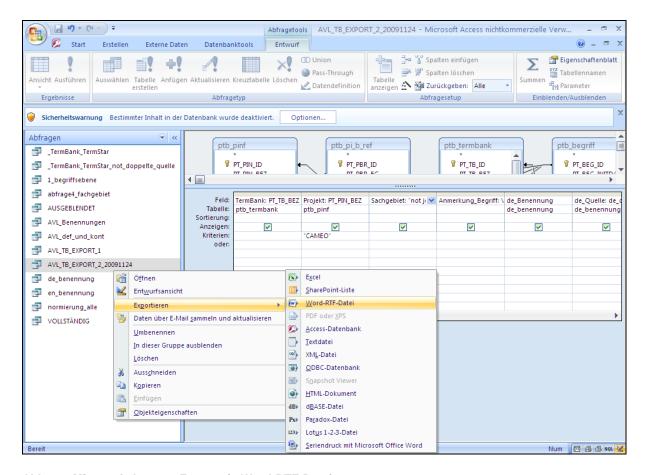


Abb. 36: Microsoft Access: Export via Word-RTF-Datei

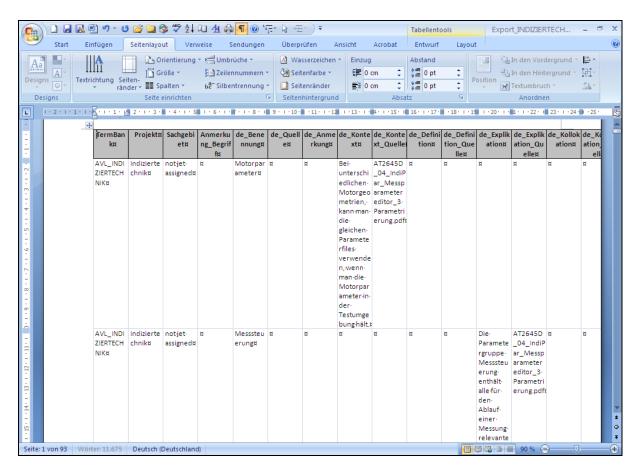


Abb. 37: Exportergebnis in Word-RTF-Datei (Auszug)

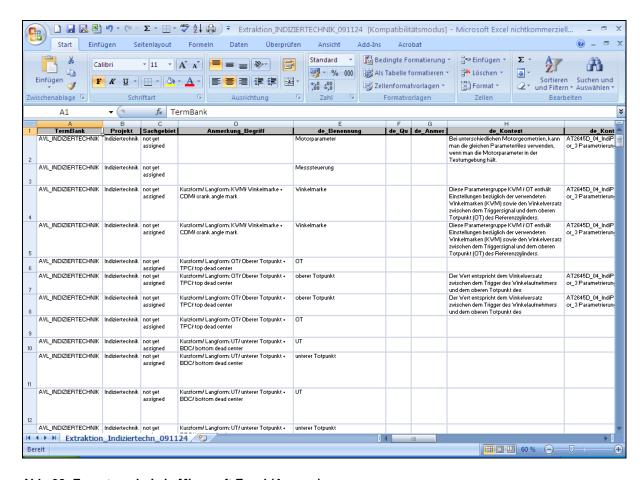


Abb. 38: Exportergebnis in Microsoft Excel (Auszug)

7.8 Ergebnis der Terminologieextraktion

Tab. 4 gibt einen Überblick über das gesamte Extraktionsergebnis. Jede Datenkategorie wurde nur einmal angeführt. Die Tatsache, dass die Anzahl der Datenkategorien nicht fortlaufend identisch sind, ist darauf zurückzuführen, dass sie ausschließlich aus ProTerm und bei Bedarf auch mehrfach (dies trifft vor allem auf *Kontext* zu) extrahiert wurden.

Tab. 4: Ergebnis: Terminologieextraktion

Dokumentations-		CAMEO	EMCON	Indizier- technik	SANTORIN	GESAMT
gruppe		CAMILO	LIVICOIN	tecillik	SANTOKIN	GLOAMI
	d					
	е	227	453	108	189	977
Termkandidat						
	е					
	n	228	454	108	186	976
Definition		24	5	4	9	42
Explikation	d	18	17	25	22	82
Kollokation	е	44	33	10	11	124
Kontext		143	141	103	38	425
Definition		22	6	4	9	41
Explikation	е	17	20	20	21	78
Kollokation	n	42	37	7	11	119
Kontext		139	147	111	35	432

7.9 Validierung der Termkandidaten

Die Extraktionsergebnisse wurden an Herrn Baumgartner (AVL) gesendet und von ihm an die für die jeweilige Dokumentationsgruppe verantwortlichen technischen Autoren übergeben und auf ihren Fachlichkeitsgrad überprüft. Im Bedarfsfall wurde Rücksprache mit Ansprechpartnern im Projekt (zum Beispiel Mitarbeiter aus der Entwicklungsabteilung oder dem Projektmanagement) gehalten. Die validierten Termkandidaten wurden abschließend als Termini in die AVL-Terminologiedatenbank übernommen.

7.10 Stopp-Wort-Listen

StW-Listen dienen dazu festzulegen, welche Termini während dem Einlesen ausgeschlossen werden und somit im Einleseergebnis nicht aufscheinen. Zielinski und Safar (2005) definiert Stopp-Wort-Listen folgendermaßen: "Stoppwortlisten beinhalten "leere" Worte (...), die uninteressant für den Terminologen sind, da sie keine terminologischen Einheiten darstellen und sie automatisch aus den TK-Listen entfernt werden sollen. Trotzdem werden diese Worte oft wegen ihrer morphosyntaktischen Struktur oder ihres häufigen Vorkommens als TK herausgefiltert." Im Zuge dieser Masterarbeit werden Stopp-Wort-Listen dazu verwendet, nicht erwünschte Termkandidaten schon vor dem Einlesen der Texte auszuschließen.

Im Vorfeld wurde bereits erwähnt, dass während der Auswahl der Termkandidaten angezeigte Benennungen ebenfalls als Stopp-Wörter kategorisiert werden können. Stopp-Wörter werden in den StW-Listen gespeichert und können mithilfe des Stopp-Wort- Editors (siehe Abb. 39 und Abb. 40) verwaltet werden.

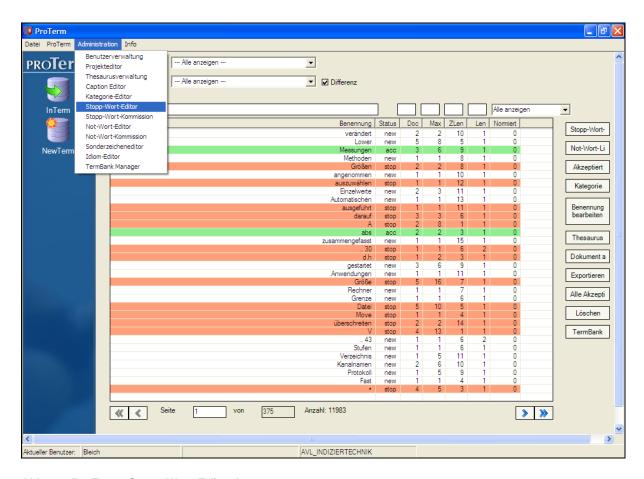


Abb. 39: ProTerm: Stopp-Wort-Editor I

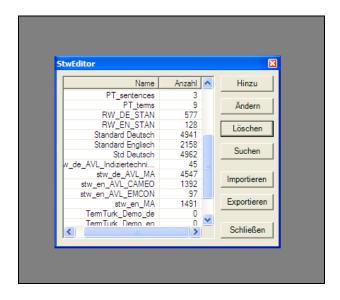


Abb. 40: ProTerm: Stopp-Wort-Editor II

Der Stopp-Wort-Editor dient dazu, neue StW-Listen hinzuzufügen, bestehende StW-Listen zu ändern oder zu löschen. Durch Klicken auf die Schaltfläche *Suchen* kann nach bestehenden Stopp-Wörtern gesucht werden. Mithilfe der Import- und Exportfunktion können StW-Listen innerhalb eines Projektteams mühelos ausgetauscht werden.

7.10.1 Erstellen neuer Stopp-Wort-Listen

Durch Klicken auf die Schaltfläche *Hinzu* kann eine neue StW-Liste erstellt werden (siehe Abb. 41). Die Stopp-Wörter können manuell eingegeben (siehe Abb. 42) oder als .txt-Format importiert werden (siehe Abb. 42 und Abb. 43).

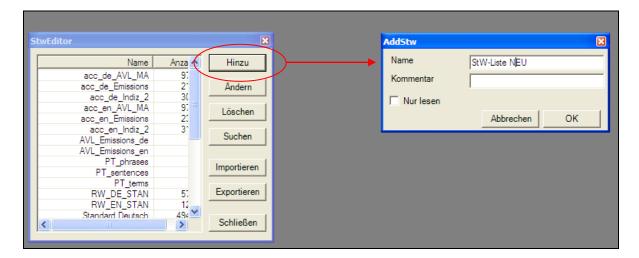


Abb. 41: ProTerm: Erstellen neuer Stopp-Wort-Listen

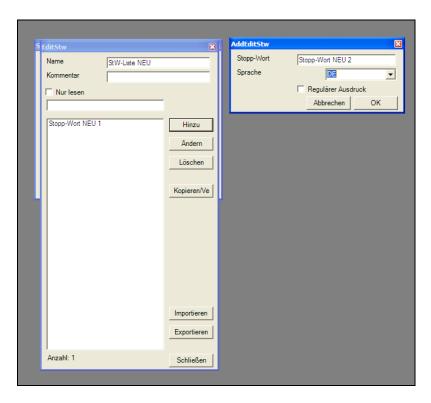


Abb. 42: ProTerm: Manuelles Hinzufügen von Stopp-Wörtern

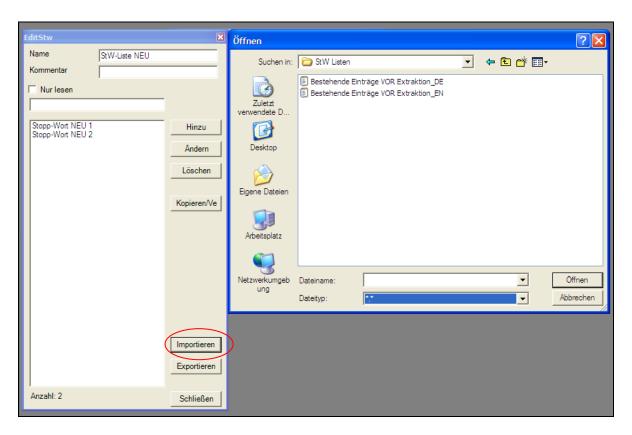


Abb. 43: ProTerm: Importieren von StW-Listen in .txt-Format

7.10.2 Verwendung der Stopp-Wort-Listen

Die StW-Listen dienen dazu, nicht gewünschte Benennungen auszuschließen. Eine der Vorgaben der AVL war es, Benennungen mit einem möglichst hohen Fachsprachlichkeitsgrad zu extrahieren, deshalb konnten allgemeinsprachliche Termini ausgeschlossen werden. Die Entwickler haben dazu StW-Listen mit allgemeinsprachlicher Terminologie in englischer und deutscher Sprache, wie sie auch von den gängigen Suchmaschinen verwendet werden, zur Verfügung gestellt. Neben diesen StW-Listen wurden zu Beginn auch die bisherigen Einträge der AVL-Terminologiedatenbank als StW-Liste generiert (siehe Abb. 43) und eingelesen. Somit konnte sichergestellt werden, dass keine bereits in der AVL-Terminogiedatenbank existierenden Termini als TK angezeigt werden. Nachdem eine Dokumentationsgruppe abgeschlossen war, wurden die dort generierten Stopp-Wörter und die akzeptierten TK im nachfolgenden Projekt als StW-Listen erstellt und eingelesen. Dadurch wurde wiederum vermieden, dass bereits während des Projektes ausgewählte Termkandidaten und Stopp-Wörter eines vorherigen Projektes nicht nochmals angezeigt werden. Dieser Vorgang wurde für jedes Projekt wiederholt und sollte dazu dienen, TK rascher ausfindig zu machen, indem das Einleseergebnis so komprimiert wurde.

7.10.3 Generieren von Stopp-Wörtern während des Auswahlverfahrens

Während des Auswahlverfahrens können als Stopp-Wörter gewünschte Benennungen durch Betätigen der Funktionstaste *F6* oder durch Klicken auf die Schaltfläche *Stopp-Wort-Liste* einer zuvor generierten Stopp-Wort-Liste hinzugefügt werden (siehe Abb. 44 und Abb. 45).

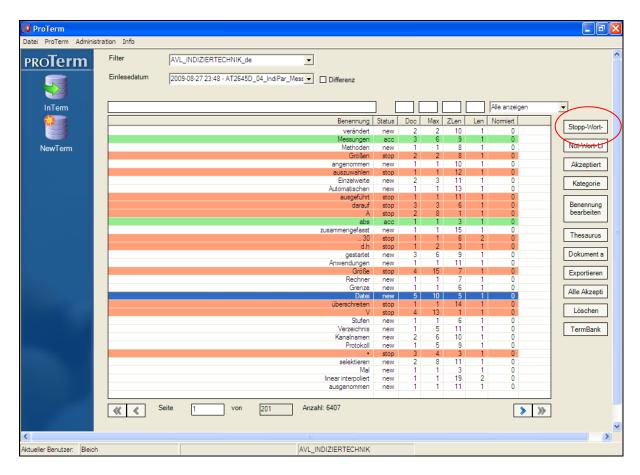


Abb. 44: Generieren von Stopp-Wörtern während des Auswahlverfahrens I

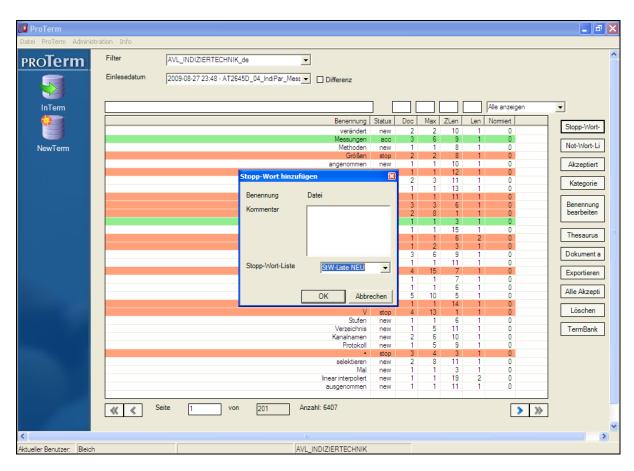


Abb. 45: Generieren von Stopp-Wörtern während des Auswahlverfahrens II

7.10.4 Ändern der Stopp-Wortlisten

Mitunter kann es vorkommen, dass ein als Stopp-Wort klassifizierter Terminus in einem Nachfolgeprojekt als Termkandidat gewünscht wird. Wird in der *Alle Anzeigen*-Ansicht (siehe Kapitel 7.2.5.1 h) gearbeitet, so ist schnell ersichtlich, welche als Stopp-Wort klassifizierte Benennung als solche nicht mehr gewünscht ist. Nachdem das betreffende Stopp-Wort markiert wurde, betätigt man die Funktionstaste *F2* oder klickt auf *Benennung aufheben*, re-klassifiziert die Benennung als *Neu* (siehe Abb. 46 bis Abb. 48) und löscht sie im Anschluss aus der jeweiligen StW-Liste, damit es nicht erneut als Stopp-Wort eingelesen wird.

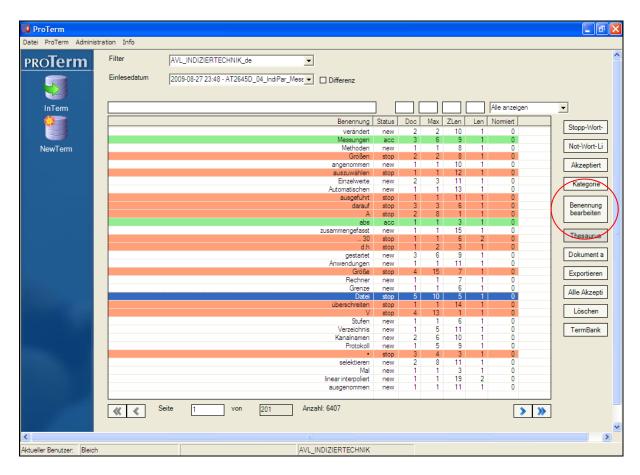


Abb. 46: ProTerm: NewTerm - Benennung bearbeiten

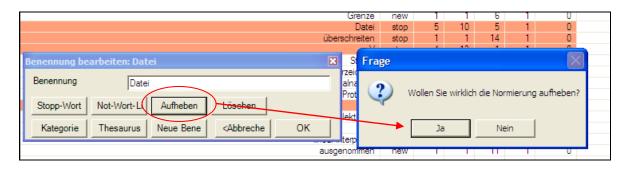


Abb. 47: ProTerm: Normierung aufheben

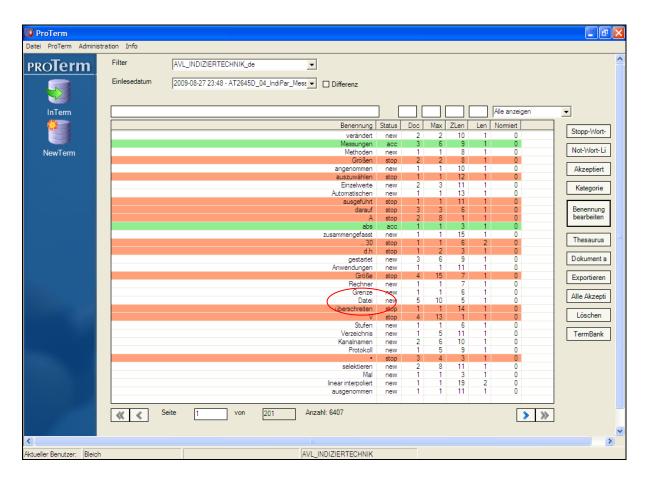


Abb. 48: ProTerm: Normierung aufgehoben

7.11 Andere Methoden

Zu Beginn dieses Masterarbeitsprojektes wurden Überlegungen angestellt, wie mit ProTerm zweisprachig Terminologie extrahiert werden kann. Die oben angeführte Methode erwies sich dabei als am wenigsten zeitintensiv. Im Folgenden sollen die anderen Ansätze vorgestellt werden.

7.11.1 Ein Filter mit allen Dokumenten einer Dokumentationsgruppe

Technische Texte weisen naturgemäß einen hohen Fachlichkeitsgrad auf. Es wurde daher angenommen, dass ein in einer Sprachversion vorkommender Fachterminus dieselbe Häufigkeit (in Dokumenten sowie absolute Häufigkeit des Terminus selbst, siehe Kapitel 7.2.5.1) im zielsprachlichen Text aufweist. Das Einlesen der Dokumente beider Sprachen hätte daher im Idealfall zu englischen und deutschen Sprachpaaren nach ihrer absoluten Häufigkeit führen sollen. Diese Methode wurde verworfen, da die angezeigten Benennungen in ProTerm von Stopp-Wörtern oder Sonderzeichen abgegrenzt sind und dies aufgrund der Syntax der beiden Sprachen nicht eins zu eins übernommen werden kann.

7.11.2 pdf-Dokumente kapitelweise einlesen

Das Einlesen beider Sprachversionen in einen Filter war nicht von Erfolg; deshalb wurden Überlegungen angestellt, wie Termini, die in den ausgangssprachlichen Texten gefunden wurden, schnellstmöglich in den zielsprachlichen Texten zu identifizieren sind. Die pdf-Dokumente wurden in ihre einzelnen Kapitel zerlegt und in einen für beide Sprachen erstellten Filter eingelesen. Das Ziel dieses Ansatzes war es, aufgrund der erhöhten Häufigkeit der Dokumente die zielsprachlichen Äquivalente zeitsparender zu identifizieren. Das Identifizieren in der daraus folgenden größeren Anzahl von (Teil-)Dokumenten erwies sich jedoch als zeitintensiver als ursprünglich gedacht, weshalb der Entschluss gefasst wurde, die Dokumente als Ganze einzulesen. Die idente Häufigkeit der Benennungen in den gesplitteten Dokumenten war auch bei diesem Ansatz in den beiden Sprachversionen nicht einwandfrei gegeben.

8. Schlussbetrachtung

Zum Abschluss dieser Arbeit sollen die Vor- und Nachteile der TE mit ProTerm zusammengefasst werden. ProTerm ist einwandfrei für die einsprachige TE anwendbar. Das angewendete Terminologieextraktionsverfahren für die zweisprachige TE kann den hybriden Extraktionsverfahren zugeordnet werden (siehe Kapitel 4.3.6). Das Einleseergebnis weist alle in den eingelesenen Dokumenten vorhandenen Benennungen aus. Mithilfe der individuell kombinierbaren Parameter (siehe Kapitel 7.2.5.1) kann sich der Terminologe rasch einen Überblick über häufig vorkommende Benennungen und potentielle TK machen. ProTerm bietet die Möglichkeit, eine breite Palette an Formaten und unterschiedlichen Zeichensätzen (siehe Kapitel 7) einzulesen und kann einen großen Umfang an Dokumenten innerhalb eines Einlesevorgangs bearbeiten (siehe Tab. 2). ProTerm ist in der Lage Mehrwortbenennungen darzustellen, die vor allem in der fachsprachlichen Kommunikation häufig verwendet werden. Die Trunkierungsfunktion während der TE-Phase kann einerseits dazu genutzt werden flektierte Wörter zu identifizieren und andererseits einheitliche Terminologie zu extrahieren (siehe Kapitel 7.2.5.1). Die Trunkierungsfunktion ist nicht nur während des TE-Prozesses hilfreich, sondern auch beim Zuordnen der TK zu den bereits bestehenden Einträgen (siehe Kapitel 7.3) sowie beim Suchen von Einträgen in der TermBank (siehe Kapitel 7.6.1). In der TermBank können die Einträge manuell nachbearbeitet werden, sollten sie nicht in ihrer Grundform extrahiert worden sein. ProTerm ermöglicht es, nach Belieben Datenelemente für Datenkategorien aus den eingelesenen Texten zu extrahieren. Dabei ist es äußerst hilfreich, dass es in jeder Arbeitsphase möglich ist, in die Dokumentansicht und die Originaltextansicht zu wechseln. Der Terminologe kann sich also jederzeit während des Extraktionsprozesses ein Bild davon machen, wie die Benennungen im Originaltext verwendet werden und bei Bedarf eine (oder mehrere) Textstelle(n) mit zu extrahieren, was für die weitere Arbeit mit der gewonnen Terminologie von Vorteil ist. Die Tatsache, dass die Quelle der Originaltexte automatisch extrahiert wird, erspart dem Terminologen viel Zeit bei der anschließenden Zuordnung der TK zu den Originaltexten. Das Erstellen von Anmerkungen auf jeder TermBank-Ebene kann für die interne Kommunikation im Rahmen eines Projektes genutzt werden. Das parallele Arbeiten auf mehreren Instanzen (siehe Kapitel 7.2.2) und das Zuordnen eines TK zu einem bestehenden Eintrag in einer anderen Sprache (siehe Kapitel 7.3) ermöglicht die zweisprachige TE. Das Auffinden der zielsprachlichen Äquivalente obliegt allerdings der Kompetenz des Terminologen. Er ist nicht nur dafür verantwortlich einen TK zu identifizieren, sondern muss sein zielsprachliches Äquivalent in den Originaltexten ausfindig machen. Da es sich bei den für diese Arbeit zur Verfügung gestellten Texten ausschließlich um Dokumentationen aus der technischen Fachsprache handelte, die systematisch von Dokumentationsexperten erstellt wurden, wurde das Auffinden der

passenden zielsprachlichen Äquivalenten erheblich beschleunigt. Auf menschliches Zutun kann daher auch bei der TE mit ProTerm nicht verzichtet werden. An dieser Stelle soll nochmal auf Zerfaß (2006) verwiesen werden, die verdeutlicht, dass eine "automatische Extraktion ihre Grenzen hat. Grenzen, die für den Menschen nicht existieren, der trotz Rechtschreibfehler oder der fehlenden Grundform des Terminus eine Beziehung z. B. zur korrekten Übersetzung erkennen kann, weil er den Text versteht." Um die Zuordnung der zweisprachigen TK noch rascher zu gestalten, wäre es von Vorteil, wenn bei der Weiterentwicklung von ProTerm das Auffinden der zielsprachlichen Kapitel automatisiert werden würde. Dies könnte folgendermaßen gestaltet werden: sobald ein TK im AT und das dazugehörige Datenelemente für eine Datenkategorie ausgewählt wurde, öffnet sich ein Fenster mit dem entsprechenden zielsprachlichen Kapitel und der Terminologe spart somit Zeit bei der Suche nach dem ZT-Kapitel (siehe Kapitel 7.4) und muss lediglich den zielsprachlichen TK identifizieren. Eine Evaluierung von ProTerm oder ein Vergleich mit anderen TET anhand der Noise/Silence- beziehungsweise Recall/Precision-Parameter (siehe Kapitel 4.3.7) hat sich als nicht zweckmäßig erwiesen. Grund dafür ist, dass ProTerm Benennungen extrahiert, vor und/oder nach deren Auftreten sich im Text ein Stopp-Wort und/oder ein Sonderzeichen befindet. Eine exakte Aufstellung der Noise/Silence- beziehungsweise Recall/Precision-Parameter war aus diesem Grund nicht möglich. Die Entwickler von ProTerm haben bereits während des praktischen Teils dieser Masterarbeit dafür gesorgt, dass mehr Datenkategorien zur Auswahl verfügbar gemacht wurden¹⁰, die Quellen automatisch extrahiert wurden, die Trunkierungsfunktion auch beim Zuordnen der TK zu den bereits vorhandenen Einträgen in der TermBank (siehe Kapitel 7.3) und beim Suchen nach Einträgen in der Term-Bank selbst (siehe Kapitel 7.6) genutzt werden kann, leere Datenkategorien in der TermBank (siehe Kapitel 7.3) farblich hinterlegt sind und dass ProTerm in der Lage ist mehr Zeichensätze einzulesen. Um ProTerm noch benutzerfreundlicher zu gestalten, soll zudem ein Auswahlelement eingerichtet werden, das dem Terminologen im NewTerm-Bereich ermöglicht zwischen "Automatischer Trunkierung"¹¹ und der Suche nach der eingegebenen Schriftzeichen zu wechseln. Der Export aus ProTerm soll ebenfalls umgestaltet werden, sodass es in Zukunft möglich sein wird in die Formate MATIF und/oder TBX, die von vielen Terminologie-verwaltungssystemen unterstützt werden, zu exportieren. Für die Arbeit mit der TermBank wäre es von Vorteil nachvollziehbar zu machen, welcher User wann, welchen Eintrag erstellt beziehungsweise bearbeitet hat.

_

¹⁰ Vor Beginn dieser Arbeit waren *Definition* und *Kontext* verfügbar.

¹¹ Im Moment ist "Automatisches Trunkieren" durch Betätigen der Funktionstaste *F*3 (siehe Kapitel 10.2) möglich.

Zum Schluss soll noch einmal auf den bedeutendsten Vorteile von ProTerm hingewiesen werden: ProTerm ermöglicht es dem Terminologen während des Extraktionsprozesses Terminologie zu bearbeiten und zusätzliche Informationen aus den eingelesene Texten zu extrahieren, somit wird der Nachbearbeitungsaufwand des Terminologen erheblich reduziert.

9. Literaturverzeichnis

Arntz, Reiner, Heribert Picht, und Fritz Mayer. *Einführung in die Terminologiearbeit.* Hildesheim [u.a.]: Olms, 2009.

Cedillo, Ana Caro. Fachsprachliche Kollokationen. Tübingen: Narr, 2004.

Dahlberg, Ingetraut. *Conceptual definitions for INTERCONCEPT*. International Classification, 1981, 8. Ausgabe zitiert in Arntz, Reiner, Heribert Picht, und Fritz Mayer. *Einführung in die Terminologiearbeit*. Hildesheim [u.a.]: Olms, 2009.

DIN, 2330. Begriffe und Benennungen: Allgemeine Grundsätze. Berlin/ Köln: Beuth, 1993 zitiert in Arntz, Reiner, Heribert Picht, und Fritz Mayer. Einführung in die Terminologiearbeit. Hildesheim [u.a.]: Olms, 2009.

DIN, 2342 Teil 1. Begriffe der Terminologielehre: Grundbegriffe. Berlin/ Köln: Beuth, 1992 zitiert in Soukup- Unterweger, Irmgard. Ein praxisorientiertes
Terminologieverwaltungsmodell für das betriebliche Umfeld. Donau-Universität Krems, 2002.

Eckstein, Karina. *Toolgestützte Terminologieextraktion*. In *Terminologiemanagement*, von Felix Mayer und Ute Seewald-Heeg, 108-120. Berlin: Bundesverband der Dolmetscher und Übersetzer e-V. (BDÜ), 2009.

Haller, Johann. *AUTOTERM: Automatische Terminologieextraktion* Spanisch/ Deutsch. In *Multiperspektivische Fragestellungen der Translation in der Romania*, von Alberto Gil, Ursula Wienen und Erich Steiner, 229-242. Frankfurt am Main: Peter Lang, 2007.

Kadric, Mira, Klaus Kaindl, und Michèle Kaiser-Cooke. *Translatorische Methodik.* Wien: Facultas, 2005.

Lieske, Christian. *Pragmatische Evaluierung von Werkzeugen für die Term-Extraktion* DTT-Symposium eTerminology. Köln, 2002.

Mügge, Uwe. Automatische Terminologieextraktion. In Translationsqualität, von Peter A. Schmitt und Heike E. Jüngst. Frankfurt am Main; Wien [u.a.]: Lang, 2007.

Pearson, J.. *Terms in context*. Amsterdam: John Benjamins Publishing Company.1998 zitiert in Zielinski, Daniel, und Yamile Ramírez Safar. *Eine Onlineumfrage zum Einsatz von Terminologieextraktions- und Terminologieverwaltungstools*. Sprachdatenverarbeitung, Fachrichtung 4.6 "Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen", Universität des Saarlandes, Saarbrücken, 2005.

Rothkegel, Annely. *Kollokationsbildung und Textbildung*. Hildesheim [u.a.]: Olms, 2009.In Sandig, Barbara (Hg.). *EUROPHRAS 92 Tendenzen der Phraseologiefoschrung*. Bochum: Brockmeyer, 499-523 zitiert in Cedillo, Ana Caro. *Fachsprachliche Kollokationen*. Tübingen: Narr, 2004.

Saß, R. Vergleichende Untersuchung von Terminologie-Extraktions-Tools. Eine computerlinguistische Arbeit mit Englisch und Deutsch. Saarbrücken: Fachrichtung 4.6 - Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen - Universität des Saarlandes (Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen, Band 21). 2004 zitiert in Zielinski, Daniel, und Yamile Ramírez Safar. Eine Onlineumfrage zum Einsatz von Terminologie-extraktionsund Terminologieverwaltungstools. Sprachdatenverarbeitung, 4.6 Fachrichtung "Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen", Universität des Saarlandes, Saarbrücken, 2005.

Schmitt, Peter A. *Anleitungen/Benutzerhinweis*. In *Handbuch Translation*, von Mary Snell-Hornby, Hans G. Hönig, Paul Kußmaul und Peter A. Schmitt, 209-213. Tübingen: Stauffenburg, 2003.

Schmitz, Klaus-Dirk. *Datenkategorien für die Terminologieverwaltung Auszug aus der ISO 12620 (1999) Computer applications in terminology – Data categories –* überarbeitet und lokalisiert fürs Deutsche. Köln, 2003.

Soukup-Unterweger, Irmgard. *Ein praxisorientiertes Terminologieverwaltungsmodell für das betriebliche Umfeld.* Donau-Universität Krems, 2002.

Thurmair, G. *Making Term Extraction Tools Usable*. Comprendium Germany.Letzte Überprüfung: 11.07.03. URL: http://www.comprendium.info/pic/papers/ EAMT-2003-TExt-article.pdf. 2003 zitiert in Zielinski, Daniel, und Yamile Ramírez Safar. *Eine Onlineumfrage zum Einsatz von Terminologieextraktions- und Terminologieverwaltungstools*. Sprachdatenverarbeitung, Fachrichtung 4.6 "Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen", Universität des Saarlandes, Saarbrücken, 2005.

Zerfaß, Angelika. "Terminologieextraktion." eDITion, 2006: 21-25.

Zielinski, D. Computergestützte Termextraktion aus technischen Texten (Italienisch), Saarbrücken: Universität des Saarlandes. [Diplomarbeit] Letzte Überprüfung: 28.06.05. URL: http://fr46.uni-saarland.de/index.php?id=433.2002 zitiert in Zielinski, Daniel, und Yamile Ramírez Safar. Eine Onlineumfrage zum Einsatz von Terminologieextraktions- und Terminologieverwaltungstools.

Sprachdatenverarbeitung, Fachrichtung 4.6 "Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen", Universität des Saarlandes, Saarbrücken, 2005.

Zielinski, Daniel, und Yamile Ramírez Safar. Eine Onlineumfrage zum Einsatz von Terminologieextraktions- und Terminologieverwaltungstools.

Sprachdatenverarbeitung, Fachrichtung 4.6 "Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen", Universität des Saarlandes, Saarbrücken, 2005.

<u>Internetquellen</u>

AVL - Unternehmen. 2010. http://www.avl.com/ (Letzter Zugriff am 17. Februar 2010).

DocuMatrix. http://www.documatrix.com (Letzter Zugriff am 3. Mai 2010).

Österreichs Bundesheer. 2010.

http://www.bundesheer.at/organisation/beitraege/lvak/zdok.shtml (Letzter Zugriff am 3. Mai 2010).

ProCom-Strasser. 2009. http://www.procom-strasser.com/ (Letzter Zugriff am 3. Mai 2010).

Semantic Web Company. 2010. http://www.semantic-web.at/1.20.resource.35.procom-strasser.htm (Letzter Zugriff am 3. Mai 2010).

Warburton, Kara. "LISA Terminology Survey." *LISA*. 2008. http://www.lisa.org/LISA-Terminology-Sur.464.0.html (Letzter Zugriff am 12. Februar 2010).

Witschel, Hans Friedrich. *GLDVPreis.* 29. Juni 2005. http://wortschatz.uni-leipzig.de/~fwitschel/papers/GLDVPreis.pdf (Letzter Zugriff am 27. April 2010).

10. Anhang

10.1 Benutzeroberfläche ProTerm

In den folgenden Abbildungen wird die Benutzeroberfläche von ProTerm für die TE dargestellt.

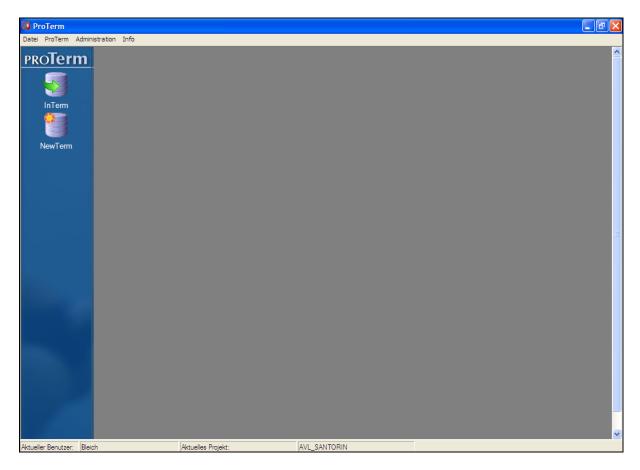


Abb. 49: ProTerm: Benutzeroberfläche Start

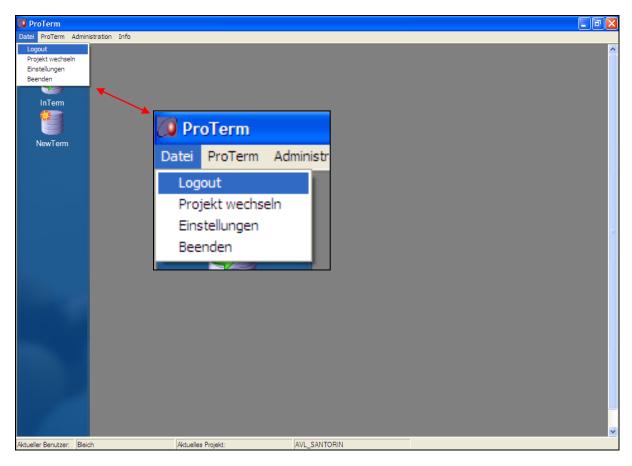


Abb. 50: ProTerm: Benutzeroberfläche Datei

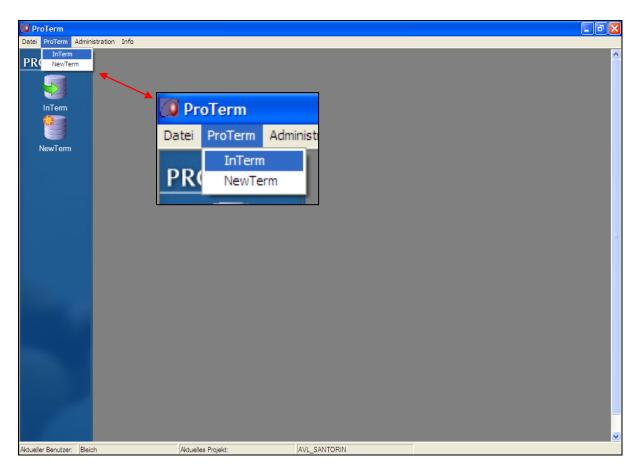


Abb. 51: ProTerm: Benutzeroberfläche ProTerm

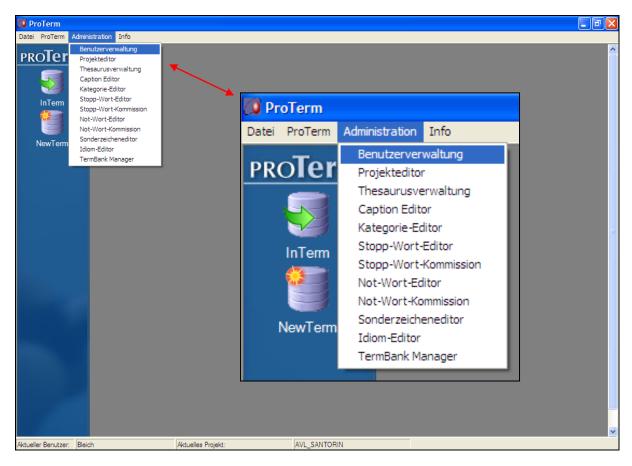


Abb. 52: ProTerm: Benutzeroberfläche Administration

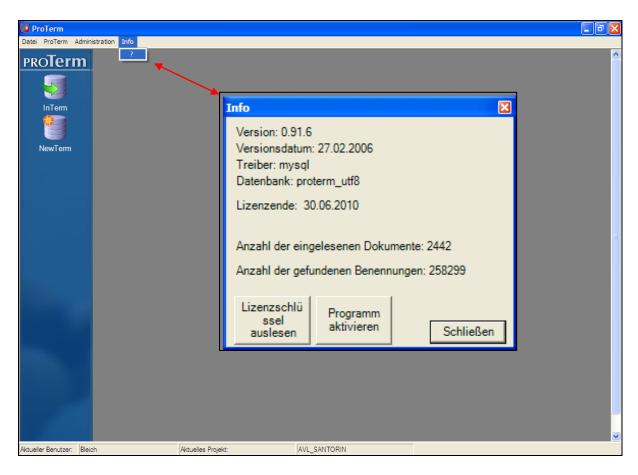


Abb. 53: ProTerm: Benutzeroberfläche Info

10.2 ProTerm-Funktionstasten

Die nachstehende Tabelle gibt einen Überblick über die Funktionen der Funktionstasten in ProTerm.

Tab. 5: ProTerm: Funktionstasten

Funktionstaste	Funktion	Anmerkung
F2	Benennung bearbeiten	
F3	Automatische Trunkierung einer zuvor markierten Benennung	Beispiel: *XY*
F4	Thesaurus	
F5	Akzeptiert	
F6	Stopp-Wort hinzufügen	
F7	Not-Wort hinzufügen	
F10	Menüleiste	

10.3 Abkürzungsverzeichnis

Die nachstehende Tabelle gibt einen Überblick über die in dieser Masterarbeit verwendeten Abkürzungen und ihre Langformen.

Tab. 6: Abkürzungsverzeichnis

Kurzform	Langform	Anmerkung
AVL	AVL LIST GmbH	
TE	Terminologieextraktion	
TET	Terminologieextraktionstool	
TEP	Terminologieextraktionsprogramm	(Eckstein 2009)
StW-Listen	Stopp-Wort-Listen	
AT	Ausgangstext	
ZT	Zieltext	
TK	Termkandidat	

10.4 Tabellenverzeichnis

Tab. 1 Ubersicht Extraktionsmaterial	17
Tab. 2: Übersicht Einlesedauer	35
Tab. 3: ProTerm: TermBank – Icons	59
Tab. 4: Ergebnis: Terminologieextraktion	67
Tab. 5: ProTerm: Funktionstasten	92
Tab. 6: Abkürzungsverzeichznis	93

10.5 Abbildungsverzeichnis

Abb. 1: Prozess Terminologieextraktion mit ProTerm	25
Abb. 2: ProTerm: Log-in	27
Abb. 3: ProTerm: Neues Projekt anlegen	28
Abb. 4: ProTerm: Neuen Filter anlegen	29
Abb. 5: ProTerm: Projekt AVL-Indiziertechnik und Filter	30
Abb. 6: ProTerm: Filter AVL_Indiziertechnik	31
Abb. 7: Filter aktivieren	32
Abb. 8: ProTerm: Einlesevorgang starten	33
Abb. 9: ProTerm: Ergebnis des Einlesevorgangs	34
Abb. 10: ProTerm: NewTerm	36
Abb. 11: ProTerm: Dokumentansicht, Suche in Dokumentansicht	37
Abb. 12: ProTerm: NewTerm- Parameter	37
Abb. 13: ProTerm: NewTerm- Benennung – Trunkierung *winkel*	38
Abb. 14: ProTerm: NewTerm- Benennung – Trunkierung *indizierter *	
Abb. 15: ProTerm: NewTerm- Benennung – Trunkierung *grafik	40
Abb. 16: ProTerm: NewTerm – Gefiltert nach Status	41
Abb. 17: ProTerm: NewTerm – Gefiltert nach Häufigkeit der Dokumente (absteigend)	42
Abb. 18: ProTerm: NewTerm – Gefiltert nach Häufigkeit der Dokumente in mehr als sechs	
eingelesenen Dokumenten	43
Abb. 19: ProTerm: NewTerm – Max – Gefiltert nach Häufigkeit (absteigend)	44
Abb. 20: ProTerm: NewTerm – Gefiltert nach Benennungen, die öfter als zweimal auftreten	
Abb. 21: ProTerm: NewTerm – Benennungen mit acht Zeichen werden angezeigt	
Abb. 22: ProTerm: NewTerm – Benennungen bestehend aus zwei Wörtern	
Abb. 23: ProTerm: NewTerm – Auswahlmenü, Farbkodierung	
Abb. 24: ProTerm: NewTerm – Zeige Drei-Wort-Benennungen, die öfter als zweimal auftreten na	
ihrer absoluten Häufigkeit sortiert	
Abb. 25: ProTerm: SelectSentence- Fenster – Textpassage Sätze	
Abb. 26: ProTerm: SelectSentence- Fenster – Textpassage Phrasen	
Abb. 27: ProTerm: AddNewTerm-Fenster	55
Abb. 28: ProTerm: Neuer Eintrag in TermBank	56
Abb. 29: ProTerm: TermBank	
Abb. 30: ProTerm: TermBank Manager	59
Abb. 31: ProTerm: TermBank – Termbankinhalt	
Abb. 32: ProTerm: TermBank – Begriffsebene	
Abb. 33: ProTerm: TermBank – Sprachebene	
Abb. 34: ProTerm: TermBank – Termebene	
Abb. 35: Microsoft Access: Export der Extraktionsergebnisse	
Abb. 36: Microsoft Access: Export via Word-RTF-Datei	
Abb. 37: Exportergebnis in Word-RTF-Datei (Auszug)	
Abb. 38: Exportergebnis in Microsoft Excel (Auszug)	
Abb. 39: ProTerm: Stopp-Wort-Editor I	
Abb. 40: ProTerm: Stopp-Wort-Editor II	
Abb. 41: ProTerm: Erstellen neuer Stopp-Wort-Listen	
Abb. 42: ProTerm: Manuelles Hinzufügen von Stopp-Wörtern	

Abb. 43: ProTerm: Importieren von StW-Listen in .txt-Format	71
Abb. 44: Generieren von Stopp-Wörtern während des Auswahlverfahrens I	73
Abb. 45: Generieren von Stopp-Wörtern während des Auswahlverfahrens II	74
Abb. 46: ProTerm: NewTerm – Benennung bearbeiten	75
Abb. 47: ProTerm: Normierung aufheben	75
Abb. 48: ProTerm: Normierung aufgehoben	76
Abb. 49: ProTerm: Benutzeroberfläche Start	87
Abb. 50: ProTerm: Benutzeroberfläche Datei	88
Abb. 51: ProTerm: Benutzeroberfläche ProTerm	89
Abb. 52: ProTerm: Benutzeroberfläche Administration	90
Abb. 53: ProTerm: Benutzeroberfläche Info	91

10.6 Index

Datenkategorie 52, 62, 67

Definition 19, 62, 67

Dokumentansicht 35, 37, 52, 79

Einleseergebnis 28, 37, 51, 68, 72, 79

einlesen 31,77

Einwortbenennung 35

Evaluierung 14, 80, 83

Explikation 19, 20, 62, 67

Export 14, 24, 53, 57, 60, 63, 64

extrahieren 17, 19, 38, 56, 72, 79

Import 14, 63, 69

Kollokation 19, 20, 62, 67

Kontext 19, 21, 52, 62, 67

linguistisches Extraktionsverfahren 10, 13

manuelle Terminologieextraktion 10, 11

Mehrwortbenennung 35

Noise 14, 80

Originaltext 11, 79

Originaltextansicht 35, 79

Parameter 37, 51, 57, 79

Precision 14, 80

ProTerm 5, 7, 14, 15, 16, 19, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 67, 69, 70, 71, 75, 76, 77, 79, 87, 88, 89, 90, 91, 92, 99

Recall 14, 80

Silence 14, 80

statistisches Extraktionsverfahren 12

StW-Liste 70, 72, 74

Suchfunktion 35, 52

TE 5, 7, 10, 11, 13, 21, 23, 24, 26, 35, 56, 57, 79, 87, 93

Terminologie 5, 7, 9, 10, 15, 16, 23, 24, 38, 56, 72, 77, 79, 84, 98

Terminologiearbeit 5, 7, 10, 17, 19, 21, 83

Terminologieextraktion 5, 9, 10, 11, 16, 24, 25, 35, 67, 83, 84, 93, 98, 99

Terminologieextraktionstool 5, 11, 93

Terminologieextraktionsverfahren 10, 13, 24, 79

Terminus 5, 11, 12, 13, 74, 77, 80

Termkandidat 11, 13, 52, 57, 67, 74, 93

Termpaar 11

TET 5, 7, 11, 14, 31, 80, 93

TK 10, 11, 12, 13, 14, 23, 35, 52, 56, 63, 68, 72, 79, 93

toolgestützte Terminologieextraktion 10

Trunkierung 38, 39, 40, 92

Zieltext 93

zweisprachige TE 6, 79

11. Curriculum Vitae

Angaben zur Person:

Verena Christina Bleich, Bakk.phil.

Am Lapp 1 Dreschnigstraße 5a/1

7152 Pamhagen 9500 Villach

Österreich

0043 (0)4242 38 580 0043 (0)650 23 23 383

verena_bleich@gmx.at

Geburtsdatum: 10. November 1983 Staatsbürgerschaft: Österreich

Sprachkenntnisse

Deutsch Muttersprache
Französisch, Englisch Ausgezeichnet
Spanisch, Russisch Basiskenntnisse

<u>Ausbildung</u>

2007 – 2010 Universität Wien, Masterstudium Übersetzen [Deutsch – Französisch – Englisch]

2008 – 2009 Institut Catholique de Paris, ISIT Institut de Management et de Communication interculturels, Erasmus Programm

2002 – 2007 Universität Wien, Bakkalaureatsstudium Übersetzen und Dolmetschen, Bakkalaurea der Philosophie

1994 – 2002 Bundesgymnasium und Bundesrealgymnasium Neusiedl am See, Matura

Fachgebiete

Terminologiewissenschaft

Terminologieextraktion

Übersetzen

<u>Lehrveranstaltungen:</u>

Französisch: Europäische Union, Recht, Wirtschaft, Naturwissenschaften,

Medizin, Tourismus, Sport

Englisch: Europäische Union, Wirtschaft, Geisteswissenschaften,

Naturwissenschaften

Persönliches Interesse:

Immobilien Management und Organisation, Wirtschaft,

Landwirtschaft, Weinbau, Technik, Technologie, Sport, Medien

und Kommunikation, Versicherungswesen

Lokalisierung

<u>Berufserfahrung</u>

04/2010 - laufend	Wissenschaftliche	Universität Wien
	Projektmitarbeiterin	Projekt: TES4IP

www.univie.ac.at/transvienna

04/2009-07/2010 Terminologieextraktion AVL List GmbH

www.avl.com/

03/2009 – laufend Wissenschaftliche Infoterm- Internationales

Mitarbeiterin Informationszentrum für

Terminologie www.infoterm.org

08/2008 Praktikum Immoconsult

Leasingges.m.b.H. Geschäftsführung www.immoconsult.eu

09/2007 – 02/2008 Karenzvertretung 07/2006 – 08/2006 Praktikum

aktikum ww

Welcome Touristic

McDonalds Kids,

Escort

UEFA Euro 08

www.fussballeskorte.at

06/2008

06/2007 — 08/2007	Praktikum Rechtsabteilung	Immoconsult. Leasingges.m.b.H
09/2006 – 05/2007	Shop Assistant	Sports Experts www.sports-experts.at
04/2006 — 05/2006	Liaison Officer	Österreich Präsidentschaft der EU www.eu2006.at
04/2005 — 05/005	Media Host	IIHF World Championship www.iihf.com

IT Skills

MS Office (Word, Excel, PowerPoint, Access)
ProTerm (Terminologieextraktion)
SDL International (MultiTerm & Trados)
TermStar NXT
webEdition
Lotus Notes
SPSS
ArcGIS

Persönliche Interessen

Berufliche und persönliche Weiterbildung Lesen Kaffee trinken Reisen Inline Skating, Radfahren Fußball, Eishockey