universität wien

MASTERARBEIT

Titel der Masterarbeit

Evaluation von Terminologieextraktionsprogrammen

Ein methodologischer Überblick

Verfasserin

Katrin Mahlstedt, Bakk. phil.

angestrebter akademischer Grad

Master of Arts (MA)

Wien, im Mai 2010

Studienkennzahl lt. Studienblatt:	A 060 351 369
Studienrichtung lt. Studienblatt:	Masterstudium Übersetzen Spanisch, Tschechisch
Betreuer:	Univ.-Prof. Mag. Dr. Gerhard Budin

Inhalt

Einleitung **1**

Motivation ... *1*

Grundlegende Begriffe ... *2*

Überblick ... *3*

1 Grundlagen der computergestützten Terminologieextraktion **5**

1.1 Ziele der computergestützten Terminologieextraktion ... *5*

1.2 Definition der computergestützten Terminologieextraktion ... *6*

1.3 Diskussion unterschiedlicher Benennungen ... *7*

1.4 Computergestützte Terminologieextraktion in der Forschung ... *9*

1.4.1 Computerlinguistik ... 10

1.4.2 Terminologie ... 11

1.4.3 Translationswissenschaft ... 12

1.5 Zusammenfassung ... *13*

2 Terminologieextraktionsprogramme **15**

2.1 Gründe für die Entwicklung von Terminologieextraktionsprogrammen ... *15*

2.2 Entstehungsgeschichte von Terminologieextraktionsprogrammen ... *16*

2.3 Charakteristische Merkmale für die Terminologieidentifikation ... *17*

2.4 Methoden zur Terminologieidentifikation ... *18*

2.4.1 Qualität des Identifikationsergebnisses ... 19

2.4.2 Methoden zur Vorverarbeitung ... 20

2.4.3 Statistische Methoden ... 22

2.4.4 Linguistische Methoden ... 24

2.4.5 Bootstrapping-Methoden ... 25

2.5 Methoden zur computergestützten bilingualen Terminologieextraktion ... *26*

2.6 Aufbau von Terminologieextraktionsprogrammen ... *28*

2.6.1 Interaktion mit dem Anwender ... 29

2.6.2 Dateneingabe und -ausgabe ... 31

2.6.3 Vorverarbeitung ... 31

2.6.4 Identifikationsverfahren ... 32

2.7 Zusammenfassung ... *33*

3 Computergestützte Terminologieextraktion in der Praxis **35**

3.1 Anwendung von Terminologieextraktionsprogrammen ... *35*

3.1.1 Vorbereitungsphase ... 35

3.1.2 Identifikationsphase ... 37
3.1.3 Nachbearbeitungsphase ... 38

3.2 Weitere Anwendungsmöglichkeiten ... 40

3.3 Probleme der computergestützten Terminologieextraktion ... 41
3.3.1 Qualität des Identifikationsergebnisses ... 41
3.3.2 Vorbereitung ... 44
3.3.3 Nachbearbeitung ... 45
3.3.4 Terminologiearbeit ... 45
3.3.5 Vielseitigkeit der Programme ... 46

3.4 Desiderata für die Weiterentwicklung ... 47

3.5 Zusammenfassung ... 48

4 Evaluation von sprachverarbeitenden Programmen 49

4.1 Konzepte, Grundlagen & Begriffe ... 49
4.1.1 Grundlagen der Softwareevaluation ... 51
4.1.2 Grundlagen der Softwarequalität ... 55

4.2 Der EAGLES-Evaluationsrahmen ... 57

4.3 Entwicklung von Anforderungen, Merkmalen & Maßen ... 59
4.3.1 Analyse der Evaluationssituation ... 60
4.3.2 Definition der Anforderungen ... 61
4.3.3 Merkmale & Maße ... 64
4.3.4 Gültigkeit & Zuverlässigkeit ... 66

4.4 Methodologie ... 66
4.4.1 Art der Evaluation ... 67
4.4.2 Evaluationsdesign ... 68
4.4.3 Methoden ... 68
4.4.4 Testtypen ... 72
4.4.5 Instrumente ... 75
4.4.6 Testmaterial ... 76

4.5 Ablauf einer Evaluation ... 77

4.6 Zusammenfassung ... 78

5 Evaluationen von Terminologieextraktionsprogrammen 81

5.1 Forschungsüberblick ... 81
5.1.1 Aus terminologischer Sicht ... 82
5.1.2 Aus translatorischer Sicht ... 84
5.1.3 Aus computerlinguistischer Sicht ... 85
5.1.4 Methodologie ... 86
5.1.5 Testmaterial ... 89

5.1.6 Vergleichbarkeit & Standardisierung 92

5.2 Quantitative Evaluation des Identifikationsergebnisses 93

5.3 Ganzheitliche Evaluationsmethodologien 98

5.3.1 Methodologische Grundlagen 98

5.3.2 Qualität des Identifikationsergebnisses 100

5.3.3 Vergleichbarkeit & Standardisierung 103

5.4 Ideen zur Entwicklung eines parametrisierbaren Evaluationsmodells 103

5.4.1 Ziele & Ausgangssituation 104

5.4.2 Anwenderprofile 105

5.4.3 Entwicklung von Anforderungen, Merkmalen & Maßen 107

5.5 Entwicklung von Methoden 109

5.5.1 Kriterienüberprüfung 109

5.5.2 Benchmarktests 112

5.6 Zusammenfassung 116

6 Zusammenfassung 117

7 Fazit 119

Bibliographie 121

Anhang 127

I. Anforderungen & Merkmale 127

II. Kriterienchecklisten 130

III. Auswahllisten 139

Abstract (deutsch) 143

Abstract (english) 145

Lebenslauf.... 147

Einleitung

Terminologiearbeit wird traditionell mit großen Unternehmen und Organisationen verbunden, die die Bedeutung von Terminologiearbeit für die Qualität ihrer Kommunikationsprozesse erkannt haben und bereit sind, in eine qualitative Terminologie zu investieren. Terminologiearbeit ist in diesem Bereich oft präskriptiv und soll zu einer effizienteren und präziseren Kommunikation beitragen, um Kosten und Arbeitszeit zu reduzieren. Doch Terminologiearbeit ist auch in anderen Bereichen mit vielfältigen Ausgangssituationen, Arbeitsbedingungen und Anforderungen von Bedeutung: z.B. für Verwaltungsorgane, die Minderheitensprachen fördern, für Lexikographen[1] und Sprachwissenschaftler, die sich etwa mit der Erforschung von regionalen Dialekten beschäftigen sowie für selbstständige Übersetzer und auch Dolmetscher, die Terminologien für den Eigengebrauch erstellen wollen. Terminologiearbeit basiert in diesen Anwendungsbereichen auf anderen, sehr unterschiedlichen Prioritäten.

Um den mit diesen Prioritäten verbundenen vielfältigen Wünschen nachzukommen, wird in verschiedenster Hinsicht versucht, die Terminologiearbeit durch Software zu unterstützen und in einigen Fällen vollständig in die Verantwortung der Programme zu legen. Es wird sowohl in der kommerziellen als auch in der universitären Softwareentwicklung fortlaufend an der Entwicklung von Terminologieverwaltungssoftware (TVS) gearbeitet. Eine Art von TVS sind Terminologieextraktionsprogramme (TEP), die bei der Erfassung von Terminologie einen Teil der sonst manuell durchgeführten Arbeit übernehmen sollen. Verschiedenste Ansätze mit unterschiedlichen Ausrichtungen und Komponenten der TEP führen zu unterschiedlichsten Ergebnissen. Für Unternehmen, Institutionen und v.a. Einzelpersonen oder kleine Firmen, die keine Ressourcen haben, selbst ein eigenes TEP entwickeln zu lassen, fällt die Entscheidung für ein Produkt auf dem unübersichtlichen durch neue Entwicklungen permanent veränderten Markt schwer.

Motivation

Zur Erleichterung dieser Entscheidung sowie zur Verbesserung von bestehenden TEP wurden und werden TEP aus verschiedenen Blickwinkeln evaluiert. Die Ausrichtung dieser Evaluationen ist unterschiedlich: Terminologen konzentrieren sich auf die Fähigkeiten zur Extraktion aller für sie relevanten Termini zur Befüllung einer Terminologiedatenbank, Translatoren legen Wert auf die Funktionalität der bilingualen Terminologieextraktion aus bereits vorhandenen Übersetzungen und Computerlinguisten analysieren die inneren Kom-

[1] In dieser Arbeit werden keine gegenderten Personenbezeichnungen verwendet, weil die Gleichberechtigung von Frauen nicht durch Sichtbarmachung, sondern nur durch Sichtbarwerdung erfolgen kann und Zwangseingriffe – auch sprachlicher Art – diesem Ziel nicht förderlich sind.

ponenten der Programme und versuchen, Verbesserungsmöglichkeiten für die Weiterentwicklung zu erkennen.

In der Methodologie gibt es ebenfalls Unterschiede, jedoch zeigt sich ein gemeinsamer Schwerpunkt aller Evaluationsmethodologien für TEP: die quantitative Analyse des Identifikationsergebnisses. Weitere Kriterien werden im Rahmen der Evaluationen oft vernachlässigt, und auch die Methodologie zur Analyse des Identifikationsergebnisses wird nicht immer Ansprüchen der Gültigkeit und Zuverlässigkeit gerecht, so dass die Evaluationsergebnisse in den seltensten Fällen als repräsentativ angesehen werden können. Die Entwicklung der jeweiligen Methodologie ist darüber hinaus aufwendig und wird durch mangelnde Standardisierung und Vergleichbarkeit erschwert.

Diese Arbeit möchte zunächst einen umfassenden Überblick über die verschiedenen Arten von TEP anhand ihrer unterschiedlichen Methoden und Funktionen, sowie über die Anwendung von TEP in der Praxis bieten – die verschiedenen Blickwinkel von Terminologen, Translatoren und Computerlinguisten sollen in dieser Arbeit gleichberechtigt berücksichtigt werden. Im weiteren Verlauf werden Grundlagen der Evaluation von sprachverarbeitender Software betrachtet – insbesondere der EAGLES-Evaluationsrahmen zur Entwicklung von anwenderorientierten Adäquatheitsevaluationen – um anschließend die Methodologie der bisherigen Evaluationen von TEP zu analysieren und Ideen zur Weiterentwicklung der Methodologie unter Einbeziehung weiterer Evaluationskriterien zu sammeln. Diese Arbeit soll zukünftigen Evaluatoren von TEP als Überblick und Inspiration für die Erarbeitung ihrer eigenen Evaluationsmethodologie dienen.

Grundlegende Begriffe

Die grundlegenden Benennungen und Begriffe der Terminologie sind für die deutsche Sprache in der DIN 2342 *Begriffe der Terminologielehre* definiert und werden in dieser Arbeit im Sinne der Norm verwendet[2]. Im Gegensatz zu diesen häufig rezipierten genormten Benennungen der Terminologielehre ist die Benennungsvielfalt im Bereich der computergestützten Teminologieextraktion (cTE) außerordentlich groß und die Begriffe werden nicht einheitlich voneinander abgegrenzt: Die englischen Benennungen *term-extraction*, *term-recognition* und *term-identification* werden z.B. bei Bowker als Synonyme bezeichnet (2008:82), Cabré *et al*. weisen daraufhin, dass andere Wissenschaftler einen Unterschied zwischen den Benennungen *extractor* und *detector* sehen, sie die Benennungen aber als Synonyme verwenden (2001:84). Für Zielinski & Ramírez-Safar hingegen existiert ein expliziter Unterschied zwischen *terminology extraction* und *term recognition* (vgl. 2005:2).

[2] Da die Erläuterung der DIN 2342 schon Bestandteil zahlreicher Publikationen ist, soll hier nicht näher auf die Definitionen eingegangen werden. Für eine Diskussion der Definitionen vgl. Arntz *et al*. 2002.

Diese terminologische Vielfalt erschwert die Suche nach passender Literatur und führt zu Missverständnissen durch begriffliche Unterschiede. Für diese Arbeit wurde die Benennung *computergestützte Terminologieextraktion* ausgewählt. Die Begründung dieser Entscheidung erfolgt im Zuge der Gegenstandsbestimmung in Kapitel 1.3.

Überblick

Das erste Kapitel betrachtet die computergestützte Terminologieextraktion als Begriff: Es werden unterschiedliche Ziele betrachtet, eine Definition erstellt und die in der Forschung verwendeten Bezeichnungen diskutiert. Im Anschluss wird ein Überblick über die unterschiedlichen Forschungsdisziplinen gegeben, die sich mit der computergestützten Terminologieextraktion auseinandersetzen.

Im zweiten Kapitel werden Terminologieextraktionsprogramme detailliert vorgestellt. Ausgehend von den Gründen zur Entwicklung und der Entwicklungsgeschichte, werden unterschiedliche Ansätze zur Terminologieidentifikation betrachtet. Zum Abschluss werden der Aufbau von Terminologieextraktionsprogrammen und ihre einzelnen Komponenten erklärt.

Das dritte Kapitel widmet sich der praktischen Anwendung von Terminologieextraktionsprogrammen. Die einzelnen Funktionen und Arbeitsschritte der computergestützten Terminologieextraktion werden nacheinander besprochen, es wird auf weitere Anwendungsmöglichkeiten von Terminologieextraktionsprogrammen eingegangen. Anschließend werden die Probleme bei der computergestützten Terminologieextraktion behandelt und Desiderata aus Forschung und Praxis für die Weiterentwicklung der Programme zusammengefasst.

Im vierten Kapitel wird in die Evaluation sprachverarbeitender Software anhand von Grundlagen der Softwareevaluation und Softwarequalität sowie unter besonderer Berücksichtigung des EAGLES-Evaluationsrahmens eingeführt. Die einzelnen Elemente des EAGLES-Evaluationsrahmens und die Arbeitsschritte zur Entwicklung einer Evaluationsmethodologie für eine anwenderorientierte Adäquatheitsevaluation werden vorgestellt.

Das fünfte Kapitel beschäftigt sich mit der Evaluation von Terminologieextraktionsprogrammen. Ausgehend von einem Forschungsüberblick über bisherige Evaluationen unter Berücksichtigung der terminologischen, translatorischen und computerlinguistischen Sicht werden die Methodologie und das verwendete Testmaterial diskutiert. Schwerpunkte bilden die Methodologie zur quantitativen Evaluation des Identifikationsergebnisses sowie eine Betrachtung ganzheitlicher Evaluationsansätze. Abschließend werden Ideen für eine Weiterentwicklung der Methodologie für die anwenderorientierte Adäquatheitsevaluation von Terminologieextraktionsprogrammen präsentiert.

1 Grundlagen der computergestützten Terminologieextraktion

Die computergestützte Terminologieextraktion wird von verschiedenen Anwendern in ihren jeweiligen Kontexten und mit teilweise sehr unterschiedlichen Zielen verwendet (vgl. Thurmair 2003:170). In diesem Kapitel sollen zunächst die verschiedenen Anwender und ihre Ziele dargestellt werden (Kap. 1.1), um dann die computergestützte Terminologieextraktion im Sinne dieser Arbeit zu definieren (Kap. 1.2) und schließlich die Auswahl der entsprechenden Benennungen begründen zu können (Kap. 1.3). Im Anschluss soll ein Einblick in die wichtigsten wissenschaftlichen Disziplinen gegeben werden, die sich mit der Forschung im Bereich der computergestützten Terminologieextraktion befassen (Kap. 1.4).

1.1 Ziele der computergestützten Terminologieextraktion

Anwender von Terminologieextraktionsprogrammen (TEP) sind neben Terminologen und technischen Schreibern auch (Computer-)Linguisten, Ethnologen, Experten verschiedenster Fachgebiete, Wissensmanager sowie Übersetzer, Dolmetscher und verschiedene Projektmanager (für Übersetzungen, technische Dokumentation, etc.) (vgl. Saß 2004:19, Cabré *et al.* 2001:53). Als gemeinsames Ziel dieser Anwender kann der Wunsch nach einer höheren Produktivität angesehen werden (Zielinski & Ramírez-Safar 2005:2).

Jeder Anwender benötigt das Extraktionsergebnis jedoch für einen bestimmten Zweck: Terminologen etwa für eine systematische Terminologieerstellung eines bestimmten Fachbereichs oder aber für eine punktuelle firmenspezifische Terminologie; Linguisten bzw. Ethnologen für die Erstellung eines Wörterbuchs der Jugendsprache oder für die Dokumentation eines regionalen Dialekts; Übersetzer z.B. für eine *ad hoc* erstellte zweisprachige Terminologieliste als Hilfsmittel für eine Übersetzung und Projektmanager z.B. für eine projektbezogene Terminologiedatenbank zur Gewährleistung der Konsistenz bei der Übersetzung einer Produktdokumentation.

Auch die Größenkategorien der verfügbaren Textressourcen können stark variieren: einige kurze Texte erfordern eine andere Behandlung als umfangreiche Textkorpora. Dementsprechend hat auch jeder Anwender andere Anforderungen an die Einstellmöglichkeiten, den Automatisierungsgrad, die Wiederholbarkeit sowie die Modifizierbarkeit und die Repräsentation der (Zwischen-) Ergebnisse, die Möglichkeiten zur Verarbeitung von Zusatzinformationen und die Exportmöglichkeiten sowie selbstverständlich an die Ergebnisse selbst: an die ausgewählten Benennungen (vgl. Zielinski & Ramírez-Safar 2006). Für Übersetzer kann eine voraussehbare Schwierigkeit bei der Übersetzung schon als ausreichendes Kriterium gelten, für die systematische Terminologiearbeit ist Repräsentativität ein wichtiges Merkmal (vgl. Thurmair 2003:171).

Jeder einzelne Anwender würde bei einer manuellen Terminologieextraktion die für seine Zwecke am besten geeigneten Termini auswählen, die nicht mit der Auswahl eines anderen übereinstimmen müssen, selbst wenn beide dieselben Vorgaben haben. Estopá Bagot hat in einem Experiment gezeigt, dass die Ergebnisse der manuellen Extraktion terminologisch relevanter Benennungen durch vier Gruppen (Übersetzer, Terminographen, Ärzte und Dokumentalisten) insgesamt 1268 Benennungen ergaben, von denen jedoch nur 119 Benennungen von allen vier Gruppen übereinstimmend ausgewählt wurden (Estopá Bagot 1999:380).

Computergestützte Terminologieextraktion wird nicht nur für die erstmalige Erstellung einer Terminologie verwendet, sondern auch für die Aktualisierung und Erweiterung von Terminologiedatenbanken, sowie für die Konsistenzüberprüfung beim Verfassen von Texten (z.B. technisches Schreiben, kontrollierte Sprache) (vgl. Saß 2004:19). Neben der verbreiteten monolingualen Terminologieextraktion (TE) gibt es auch Formen der bilingualen TE und nicht immer werden Terminologieextraktionsprogramme für begriffsorientierte Terminologiearbeit verwendet, da auch nicht in allen Bereichen begriffsorientierte Terminologiearbeit die einzig anerkannte Methode ist (vgl. Cabré 1999:13f.). Die extrahierte Terminologie repräsentiert auch nicht in jedem Fall die gesamte Terminologie des entsprechenden Fachbereichs, wie von Kageura *et al*. definiert (2001:154), sondern kann auch nur einmalig für eine Übersetzung verwendet werden oder als Grundlage für eine erweiterbare Terminologiedatenbank dienen (vgl. Fleischmann 2006:31).

Die computergestützte Terminologieextraktion kann also für jede Anwendergruppe einzeln im Hinblick auf ihre Anforderungen, Möglichkeiten und Ergebnisse betrachtet werden, oder aber aus einem möglichst weiten Blickwinkel, der Platz für verschiedene Anwender mit unterschiedlichen Bedürfnissen birgt. Diese Arbeit möchte im Rahmen ihrer Möglichkeiten eine möglichst umfassende Sicht auf Terminologieextraktionsprogramme bieten, wird aber trotz dieses Anspruchs in zahlreichen Aspekten lediglich eine Auswahl darstellen können und nicht vollständig alle unterschiedlichen Ansätze und Möglichkeiten präsentieren und diskutieren können.

1.2 Definition der computergestützten Terminologieextraktion

In Sinne des umfassenden Blickwinkels soll auch die verwendete Definition Spielraum und Anpassungsmöglichkeiten für verschiedene Bedürfnisse bieten. Aus diesem Grund soll die Definition keine Formulierungen enthalten, die auf einzelne Anwendergruppen oder Anwendungsfälle zugeschnitten sind. Solche Definitionen finden sich z.B. für Übersetzer bei Mügge (2007:713f.) oder für die systematische Terminologiearbeit bei Witschel (2005:9).

Da auch der Automatisierungsgrad und die Einstellmöglichkeiten zwischen den einzelnen Programmen sehr stark variieren, soll die Tätigkeit des Menschen in der Definition ebenso aufscheinen, wie eine mögliche hohe Automatisierung, die z.B. Kageura *et al*. beschreiben (2001:152). Darüber hinaus sollen sich die verschiedenen Teilstufen der Extraktion in der Definition widerspiegeln: in dieser Arbeit wird die Extraktion nicht auf das automatisierte Identifizieren der Benennungen beschränkt, sondern beinhaltet auch eine mögliche Nachbearbeitung der Daten (vgl. Lieske 2002:109) sowie den Export des Ergebnisses. Die anhand dieser Kriterien entwickelte Definition lautet wie folgt:

> *Computergestützte Terminologieextraktion ist eine menschliche Tätigkeit, bei der mithilfe eines Computerprogramms im Rahmen eines zu einem bestimmten Grad automatisierten Vorgangs potentielle Termini und zugehörige Zusatzinformationen aus einem bestimmten elektronisch lesbaren Text oder Textkorpus identifiziert, angezeigt, bearbeitet und/oder exportiert werden.*

1.3 Diskussion unterschiedlicher Benennungen

Im Sinne dieser umfassenden Definition des Begriffs sollen nun die in der Forschung verwendeten Benennungen diskutiert und die Entscheidung für *computergestützte Terminologieextraktion* begründet werden.

Terminologie

Im Deutschen muss zunächst eine Entscheidung für *Term-* oder *Terminologie-* getroffen werden. *Term* ist kein deutsches Wort, sondern die englische Bezeichnung für Benennung bzw. auch für Terminus (vgl. Arntz *et al*. 2002:41). *Term* ist somit weder einem eindeutigen Begriff zugeordnet, noch kann der Vorzug gegenüber zur Verfügung stehenden deutschen Bezeichnungen durch seine Bedeutung begründet werden. In deutschen Publikationen werden nichtsdestotrotz zahlreiche Benennungen mit dem Bestandteil *Term* verwendet, z.B. Termkandidat, Termextrahierung, Termdatenbank. Da jedoch für diese Zwecke geeignete deutsche Benennungen zur Verfügung stehen, sollen die deutschen Benennungen verwendet werden, die mit *Terminologie* gebildet werden.

Extraktion

In der deutschen Sprache ist die Anzahl der verwendeten Bezeichnungen für die Extraktion nicht so groß wie im Englischen[3]. Vor allem werden *Extraktion* (bzw. *Extrahierung*), *Exzerp-*

[3] In englischen Publikationen verwendet fast jeder Autor eine eigene Bezeichnung, u.a. *term extraction, term-extraction bzw. terminology extraction, term-, term* oder *terminology detection, identification, mining,*

tion, *Erkennung* und *Identifikation* (bzw. *Identifizierung*) verwendet – unter diesen Benennungen soll nun eine ausgewählt werden. Ziel ist es erneut, eine möglichst umfassende nennung zu finden, um der umfassenden Definition gerecht zu werden.

Erkennung bezeichnet laut Zielinski die Suche nach bekannten bzw. unbekannten Termini durch den Vergleich von Listen potentieller Termini mit einer bestehenden Terminologiedatenbank (2005:2, 24ff.). Um eine Doppeldeutigkeit zu verhindern, sollten *Erkennung* und *Extraktion* nicht synonym verwendet werden. Unter *Erkennung* wird in dieser Arbeit der Abgleich von potentiellen Termini mit bestehenden Terminologiedatenbanken für Zwecke der Terminologieextraktion, sowie für Konsistenzprüfungen oder Erweiterungen der Terminologiedatenbanken verstanden (vgl. Zielinski & Ramírez-Safar 2006:28, und 2005:2).

Mit der Benennung *Identifikation* (bzw. *Identifizierung*) ist ebenfalls ein recht enger Begriff verbunden, der die potentiellen Termini lediglich identifiziert, aber keine zusätzlichen Informationen zur Verfügung stellt bzw. keine Weiterverarbeitung ermöglicht (vgl. Lieske 2002: 110). Diese Benennung soll in dieser Arbeit für den internen Teilprozess verwendet werden, in dem potentielle Benennungen identifiziert werden.

Exzerption, *Extrahierung* und *Extraktion* sind tatsächlich als Synonyme anzusehen, da zwischen den ihnen zugeordneten Begriffen kein Bedeutungsunterschied zu erkennen ist. Die Bezeichnung *Terminologieextraktion* wird u.a. von Gasser, Mügge, Saß, Witschel, Zerfaß und Zielinski & Ramírez-Safar verwendet, und soll daher aufgrund ihrer relativ großen Verbreitung auch in dieser Arbeit den Vorzug erhalten.

Computergestützt

Im Zusammenhang mit Terminologieextraktionsprogrammen wird oft darauf hingewiesen, dass eine umfangreiche Nachbearbeitung durch den Anwender notwendig ist (vgl. Ahmad & Rogers 2001:728). Aus diesem Grund werden die Bezeichnungen *halbautomatisch* bzw. *semi-automatic* oder *computergestützt* bzw. *computer-aided* gegenüber *automatisch* bzw. *automatic* bevorzugt (vgl. Bowker 2008:82).

Es ist schwierig, eine Grenze zwischen *automatisch* und *halbautomatisch* zu ziehen. Da eine möglichst umfassende Benennung gesucht wird, soll *computergestützt* verwendet werden. Diese Bezeichnung lässt offen, wie hoch der Automatisierungsgrad ist, und wie viel Nachbearbeitung durch den Anwender erfolgt. Es soll also von *computergestützter Terminologieextraktion* gesprochen werden.

---

recognition, acquisition – sowie vorangestellt *automatic, semi-automatic, computer-aided* oder auch *term candidate extraction.*

Potentielle Termini bzw. potentielle Benennungen

Ahmad & Rogers schlagen vor, von *term candidate extraction* zu sprechen (vgl. 2001:755). Auch Vivaldi & Rodríguez bevorzugen die Formulierung *term candidate* als Bezeichnung für die gefundenen Benennungen, sprechen jedoch nicht von *term candidate extraction*, sondern von *term extraction* (vgl. 2006: 12). Lieske bezeichnet die gefundenen Benennungen als „sprachliche Ausdrücke, die möglicherweise Fachvokabeln sind (sogenannte *Term-Kandidaten*)“ (2002:110, Z. 26f.).

Es ergibt sich erneut die bereits diskutierte Problematik rund um den Bestandteil *Term*, der nach Möglichkeit vermieden werden sollte. Die Bezeichnung selbst könnte vielleicht treffender *Benennungskandidat* lauten. Da es aber unter anderem Ziel dieser Arbeit ist, größere Einheitlichkeit in der Terminologie zu schaffen, und keine neuen Benennungen zu erfinden, soll auf die Verwendung von *Benennungskandidat* verzichtet werden, und statt dessen von *potentiellen Benennungen* bzw. *potentiellen Termini* gesprochen werden. Letztere Bezeichnung beinhaltet auch bereits die Möglichkeit, dass das Extraktionsprogramm Zusatzinformationen ebenfalls extrahiert, so dass das Ergebnis nicht eine reine Liste *potentieller Benennungen* ist, sondern eine Sammlung *potentieller Termini*, die teilweise bereits untereinander vernetzt sind, und mit Zusatzinformationen versehen eine solide Grundlage für eine Terminologiedatenbank bilden können.

Programm

Auch auf die Bezeichnung *Tool* wird verzichtet. Die Bezeichnung ist abgesehen vom nicht notwendigen Anglizismus irreführend, da *Tool* üblicherweise ein Hilfsprogramm geringen Umfanges bzw. eine kleine Zusatzkomponente eines Programmes bezeichnet (vgl. GIS-Glossar 2006), und damit nur auf einige wenige Terminologieextraktionsprogramme zutrifft. Aus diesem Grund werden in dieser Arbeit die Benennungen *Programm* bzw. *Terminologieextraktionsprogramm* verwendet.

1.4 Computergestützte Terminologieextraktion in der Forschung

Forschungsarbeiten, die sich mit computergestützter TE beschäftigen, werden in unterschiedlichen wissenschaftlichen Disziplinen verfasst: in der Computerlinguistik, in der Terminologie und auch in der Translationswissenschaft. Um einen Überblick über die unterschiedlichen Hintergründe zu erhalten, sollen diese drei Disziplinen und die Rolle der computergestützten TE in ihrem Rahmen, sowie einige charakteristische Grundlagen der Forschungsansätze hier kurz vorgestellt werden.

1.4.1 Computerlinguistik

Die Computerlinguistik, eine in sich bereits stark interdisziplinäre Forschungsrichtung, die ihre Grundlagen und Methoden einerseits aus der Informatik, andererseits aus der Linguistik bezieht, beschäftigt sich „mit der maschinellen Verarbeitung natürlicher Sprache“ (Carstensen *et al.* 2004:1). Die computergestützte TE ist als eine Anwendung der computergestützten Terminologie (auch *Computational Terminology*) dem Teilbereich der praktischen Computerlinguistik zuzurechnen (vgl. Carstensen *et al*. 2004:6f.).

Die computergestützte Terminologie umfasst neben der Extraktion von Termini und Relationen auch die Strukturierung von Fachwortschätzen, Terminologiedatenbanken, Wortnetzen und Ontologien, die dann wiederum für andere Anwendungen der Computerlinguistik, z.B. *information retrieval* und *text mining* (z.B. Suchmaschinen), bzw. für die Informationsextraktion verwendet werden können (vgl. Heid 2004:472).

Witschel sieht die computergestützte TE als Teilbereich des *text mining*, das sich mit der „Suche nach Informationen in unstrukturierten natürlichsprachigen Daten“ beschäftigt und somit Methoden für die TE liefern kann (vgl. 2005:9). Weitere Überschneidungen gibt es mit dem Bereich *automatisches Indexieren*, bei dem mithilfe statistischer Methoden repräsentative Indexterme aus Texten ausgewählt werden (vgl. Witschel 2005:31). Ähnliche Ansätze werden in statistischen Methoden für computergestützte TE verwendet (siehe Kapitel 2.4.3), die u.a. für die Korpuslinguistik entwickelt worden sind (vgl. Fleischmann 2006:42).

Durch den linguistischen Schwerpunkt der Computerlinguistik, werden z.B. bei der Verarbeitung von Textkorpora für die Vorverarbeitung zahlreiche linguistische Ressourcen und Methoden verwendet (siehe Kapitel 2.4.2), die u.a. für die maschinelle Übersetzung entwickelt worden sind (vgl. Fleischmann 2006:42).

Computerlinguisten erforschen meist die Effizienz von TEP bzw. die Funktionalität einzelner Komponenten von TEP, die sie oft selbst programmiert haben oder deren genauen Aufbau sie kennen. Dadurch kann ein besonders detaillierter Blick in den technischen Ablauf der computergestützten TE gewährt und eine besondere Optimierung einzelner Arbeitsschritte ermöglicht werden (vgl. Witschel 2005, Kageura *et al*. 2001). Diese Art der Evaluation, bei der die inneren Abläufe der Programme bekannt und Bestandteil der Evaluation sind, wird auch als *Glass-Box-Evaluation* bezeichnet. Kommerziell vermarktete Produkte werden normalerweise nicht aus computerlinguistischer Sicht betrachtet, da ihr Aufbau und ihre Komponenten meist nicht bekannt sind (vgl. Saß 2004:67). Die rein äußerliche Betrachtung der Programme, bei der kein Einblick ins Innere gewährt und evaluiert wird, wird als *Black-Box-Evaluation* bezeichnet.

1.4.2 Terminologie

Wichtige Forschungsbereiche der Terminologie sind die Terminologie an sich (Termini), Fachsprachen, Begriffe vs. Benennungen, Terminologienormung und die praktische Terminologiearbeit. Computergestützte TE ist als Bestandteil der praktischen Terminologiearbeit zu sehen, wird jedoch auch im allgemeinen Teilbereich der computergestützten Terminologiearbeit erforscht (vgl. Arntz *et al*. 2002). Bei Wright & Budin (2001) wird die Terminologieextraktion in einem Kapitel mit der Korpuslinguistik behandelt, da ein enger Zusammenhang zwischen Textkorpora als Grundlage für die Extraktion und der Extraktion selbst gesehen wird. Es werden u.a. computerlinguistische Ansätze beschrieben und die Qualität der potentiellen Termini im Bezug auf *termhood* und *unithood* (siehe 2.3) sowie bekannte Probleme bei der Terminologieidentifikation diskutiert.

Einige Klassifizierungen der praktischen Terminologiearbeit sind relevant für die Anforderungen zahlreicher Terminologen an TEP. Für Arntz *et al*. ist das Soll-Ziel jeder praktischen Terminologiearbeit die systematische Terminologiearbeit. Diese wird jedoch oft nicht erreicht, sondern lediglich ihre Vorstufe in Form von punktuellen Untersuchungen oder Bearbeitungen größerer Begriffsfelder (vgl. Arntz *et al*. 2002:216ff.). Cabré hingegen erkennt auch punktuelle Terminologiearbeit (auch *ad hoc* genannt) als vollwertige Terminologiearbeit zur Erfüllung spezifischer Ziele an (vgl. Cabré 1999:129-159).

Darüber hinaus wird unterschieden zwischen einer Erfassung des Ist-Zustands (deskriptive Terminologiearbeit) und einer Normung des Soll-Zustands (normende bzw. auch präskriptive Terminologiearbeit[4]) sowie zwischen der Akzeptanz von Benennungen als Grundlage und Eintragsstichwort (benennungsorientierte Terminologiearbeit) und der Bevorzugung von Definitionen als Form der Repräsentation eines Begriffs mit Zuordnung der entsprechenden Benennungen (begriffsorientierte Terminologiearbeit[5]).

Da die Begriffsorientierung eine wichtige Grundlage der terminologischen Forschung darstellt, akzeptieren Terminologen oft nur einen begriffsorientierten Ansatz (vgl. Fleischmann 2006:32-35), der teilweise auch eine bilinguale computergestützte TE von vornherein ausschließt (Fleischmann 2006:13). Für Terminologen ist es darüber hinaus in manchen Fällen relevant, dass die potentiellen Termini möglichst repräsentativ für das zugrundeliegende Fachgebiet sind (vgl. Kageura *et al.* 2001:154).

Zahlreiche Publikationen von Terminologen werden jedoch in Zusammenarbeit mit Unternehmen verfasst. Die extrahierte Terminologie konzentriert sich dementsprechend meist auf einen Teilbereich eines Fachgebiets – auf die jeweilige unternehmensspezifische Termi-

[4] Für eine Diskussion dieser Unterscheidungen vgl. Arntz *et al*. 2002:227.

[5] Für eine Diskussion dieser Unterscheidungen vgl. Fleischmann 2006:32-35.

nologie. Die Terminologiearbeit ist in diesen Fällen punktuell und soll zur Erarbeitung einer präskriptiven Terminologie dienen.

Meist werden kommerzielle Produkte im Rahmen von Black-Box-Evaluationen getestet, bei denen nicht näher auf die einzelnen Komponenten eingegangen werden kann, sondern das Resultat als Ganzes bewertet wird (vgl. Gasser 2004, Fleischmann 2006). Darüber hinaus werden oft die Einführung eines Terminologiemanagements in dem jeweiligen Unternehmen und die damit verbundenen (Kommunikations-)Prozesse, sowie die computergestützten Möglichkeiten der Weiterverarbeitung der Daten zur Validierung bzw. zum Export in Terminologiedatenbanken behandelt (vgl. König 2009).

1.4.3 Translationswissenschaft

In der Translationswissenschaft erforscht vor allem die Übersetzungswissenschaft computergestützte Terminologieextraktion, meist als Teilaspekt der sogenannten CAT-Programme (*computer-aided translation*). Textkorpora spielen in diesem Zusammenhang nur selten eine Rolle. Oft verwenden Übersetzer TEP, um aus einzelnen Texten vor der Übersetzung voraussichtlich unbekannte oder schwer zu übersetzende Elemente zu extrahieren, die dann im Vorfeld des Übersetzungsvorgangs recherchiert werden (vgl. Mügge 2007:713).

Im Rahmen von großen Übersetzungsprojekten werden TEP auch eingesetzt um eine vorgeschaltete Terminologiearbeit zu ermöglichen, aus der eine präskriptive bilinguale Terminologie erstellt wird, die die einzelnen Übersetzer dann verwenden sollen. In einigen Fällen werden Übersetzer auch als „klassische Terminologen“ tätig und erfassen Terminologie aus größeren Textkorpora für die eigene Verwendung bzw. auf Kundenwunsch (Zielinski & Ramírez-Safar 2005:14). Generell gibt es in diesem Bereich auch Interesse an bilingualer Terminologieextraktion anhand von großen *translation memories* (gespeicherte alignierte Übersetzungen) (vgl. Zielinski & Ramírez-Safar 2005:20).

Als relevante translationswissenschaftliche Grundlage sollte u.a. die *Skopostheorie*[6] genannt werden, die den Zweck einer jeden Übersetzung in den Vordergrund stellt, wodurch jede einzelne Übersetzung eine individuelle Aufgabe ist, die individuelle Terminologiearbeit erfordert, also in jedem Fall eine erneute Extraktion und Auswahl der relevanten Termini erforderlich macht. Darüber hinaus ist die *Theorie des translatorischen Handelns*[7] relevant, die die Zusammenwirkung der verschiedenen Akteure im Übersetzungsprozess beschreibt,

[6] Hier soll auf nähere Ausführungen verzichtet werden, für eine umfassendere Behandlung der Skopostheorie vgl. Dizdar 1999.

[7] Hier soll auf nähere Ausführungen verzichtet werden, für eine umfassendere Behandlung des translatorischen Handelns vgl. Risku 1999.

und somit Anforderungen für geteilte Workflows, Möglichkeiten der Interaktion, Kommunikation und Entscheidungsfindung definiert.

Publikationen beschränken sich im translationswissenschaftlichen Bereich meist auf Artikel in Fachzeitschriften, langjährige Forschung zur computergestützten TE wird nur von wenigen betrieben (z.B. Zielinski). Diplomarbeiten, die von Übersetzern verfasst werden, sind meist der Terminologie zuzuordnen. Jedoch gibt es auch einige Arbeiten, die sich vor allem mit den Prioritäten von Übersetzern befassen (vgl. Saß 2004).

Übersetzer befassen sich fast ausschließlich mit kommerziell vermarkteten Produkten und suchen oft nach Formen der bilingualen Terminologieextraktion (vgl. Zerfaß 2006). Sie reflektieren nur selten die internen Komponenten und legen besonderen Wert auf Kompatibilität der Programme mit üblichen Text- und Tabellenverarbeitungsprogrammen (vgl. Zielinski & Ramírez-Safar 2005:15), oder versuchen preisgünstige Alternativen zu TEP zu finden (vgl. Mügge 2007).

Ein neueres Forschungsthema der Translationswissenschaft sind die weiteren Anwendungsmöglichkeiten von Terminologieextraktionsprogrammen im Übersetzungsprozess, wie z.B. Konsistenzprüfung, Erweiterung bestehender Terminologiedatenbanken bzw. Terminologieerkennung (vgl. Wohlfahrt 2008).

1.5 Zusammenfassung

In diesem Kapitel wurde der Gegenstand der vorliegenden Arbeit definiert. In einem ersten Schritt wurden die unterschiedlichen Ziele einzelner Anwender bei der Verwendung von TEP beschrieben und daraufhin eine umfassende Definition des Begriffs computergestützte Terminologieextraktion formuliert. In Folge wurde die für diese Arbeit ausgewählte Terminologie ausführlich begründet und im Anschluss ein Einblick in die unterschiedlichen Forschungsdisziplinen gegeben, die sich mit computergestützter Terminologieextraktion beschäftigen.

Das folgende Kapitel beschäftigt sich mit Terminologieextraktionsprogrammen, ihrer Entwicklung, ihren einzelnen Komponenten und unterschiedlichen Ansätzen.

2 Terminologieextraktionsprogramme

Nachdem nun der Untersuchungsgegenstand beschrieben wurde, soll dieses Kapitel die Programme für die Terminologieextraktion näher betrachten. Zunächst werden die Beweggründe für die Entwicklung von TEP (Kap. 2.1) sowie ihre Entstehungsgeschichte (Kap. 2.2) skizziert. Anschließend sollen die unterschiedlichen charakteristischen Eigenschaften von Termini dargestellt werden, die für die Identifikation von potentiellen Termini verwendet werden (Kap. 2.3). Darauffolgend sollen die unterschiedlichen Terminologieidentifikationsmethoden, einschließlich Vorverarbeitungsmethoden vorgestellt werden (Kap. 2.4). Im Anschluss sollen Möglichkeiten zur computergestützten bilingualen Terminologieextraktion aufgezeigt werden (Kap. 2.5) und abschließend der Aufbau von TEP anhand der einzelnen Komponenten betrachtet werden (Kap. 2.6).

Sowohl Softwarefirmen als auch Forscher und Forschungsprojekte entwickeln Terminologieextraktionsprogramme. In der Forschung werden Glass-Box-Evaluationen durchgeführt und im Rahmen von Publikationen die einzelnen Komponenten und der Aufbau der Programme beschrieben. Eine solch detaillierte Dokumentation kommerziell vermarkteter Programme ist in den meisten Fällen nicht erhältlich, da Softwarefirmen solche Informationen als Firmengeheimnis zum Schutz vor der Konkurrenz nicht veröffentlichen. Forschungsarbeiten, die sich mit kommerziellen Produkten beschäftigen, können dementsprechend nur in Black-Box-Evaluationen untersuchte „von außen sichtbare“ Merkmale und Identifikationsergebnisse beschreiben (vgl. Saß 2004:67). Aus diesem Grund basiert dieses Kapitel vorwiegend auf Beschreibungen von Terminologieextraktionsprogrammen, die nicht kommerziell vermarktet werden. Falls auch über kommerziell vermarktete Programme Informationen zur Verfügung stehen, wird gesondert darauf hingewiesen.[8]

In diesem Kapitel sollen lediglich die Grundlagen der einzelnen Ansätze dargestellt werden. Im Rahmen dieser Arbeit soll keine Wertung der einzelnen Verfahren vorgenommen werden, da sich jeder Anwender anhand seiner persönlichen Anforderungen für ein bestimmtes TEP entscheiden muss.

2.1 Gründe für die Entwicklung von Terminologieextraktionsprogrammen

Die Extraktion von Terminologie ist so alt wie die Terminologie selbst, im Rahmen der Terminologiearbeit wurden selbstverständlich auch vor der Entwicklung von Computerprogrammen Termini bzw. Benennungen aus Texten extrahiert. Früher wurde diese Extraktion vollständig manuell durchgeführt, indem Texte durchgelesen bzw. überflogen wurden und potentielle Termini in einer Liste gesammelt wurden (vgl. Mügge 2007: 714).

[8] Einen Überblick über kommerzielle Produkte bieten Zielinski & Ramírez-Safar (2005:4ff.).

Die Praxis der manuellen Terminologieextraktion ist auch heute noch weit verbreitet (vgl. Zielinski & Ramírez-Safar 2006:27), jedoch ist diese Art der Terminologieextraktion sehr zeit- und somit auch kostenintensiv und ab einer gewissen Korpusgröße im Vergleich zu computergestützter TE ineffizient (vgl. Fleischmann 2006:122). Aus diesen Gründen entstand Interesse an einer Entlastung der Extraktoren und einer möglichen größeren Effizienz durch maschinelle Unterstützung.

Ein weiterer Beweggrund zur Entwicklung von TEP ist die Hoffnung auf verlässlichere Auswahlkriterien für terminologisch relevante Benennungen. Sager verweist auf die durch die elektronische Verfügbarkeit von Textkorpora entstandenen neuen Möglichkeiten, solche Auswahlkriterien auf der Basis von Daten und Statistiken zu ermitteln, so dass die Intuition nicht die einzige Entscheidungsgrundlage bleiben muss (vgl. 2001:762).

2.2 Entstehungsgeschichte von Terminologieextraktionsprogrammen

In den 80er Jahren wurden erste Versuche unternommen, Daten aus großen maschinell lesbaren Textkorpora automatisiert zu extrahieren (vgl. Ahmad & Rogers 2001:727). Die Qualität dieser ersten Programme war jedoch nicht besonders überzeugend, obwohl die ersten Ansätze seit den 80er Jahren fortlaufend verbessert wurden (vgl. Vivaldi & Rodríguez 2006:12). Das erste weitläufig bekannte, jedoch noch stark verbesserungsfähige Terminologieextraktionsprogramm *Termino* erschien 1990, und im Laufe der 90er Jahre wurden zahlreiche weitere Programme entwickelt.[9]

Über die Jahre hinweg bestätigen Wissenschaftler regelmäßig Fortschritte bei der Entwicklung von TEP: Cabré *et al*. sehen bereits in den 90er Jahren erste vielversprechende Ergebnisse aufgrund des vergrößerten Repertoires an elektronisch gespeicherten Textkorpora (vgl. 2001:53). Ahmad & Rogers bestätigen 2001 erste brauchbare Entwicklungen, die sich aus Forschungsergebnissen auch auf am Markt erhältliche Produkte auswirken (2001:755). Vivaldi & Rodríguez befinden im Jahre 2006, dass die Terminologieextraktionsprogramme eine gewisse Reife erreicht haben (vgl. 2006:12).

Wie bereits in Kapitel 1.4 erwähnt, gibt es unterschiedliche Ansätze zur Umsetzung von computergestützter TE. Eine mögliche Einteilung sind die Kategorien linguistisch, statistisch bzw. hybrid als eine Kombination dieser beiden (siehe auch Kapitel 2.3, 2.4). Die statistischen Ansätze basieren auf den Ideen des *information retrieval* bzw. der Informationsextraktion. Erste Entwicklungen gab es bereits in den 70er Jahren, jedoch nicht mit dem deklarierten Ziel der Terminologieextraktion, sondern z.B. für automatisches Indexieren (vgl. Witschel 2005:35). Bedeutende Entwicklungen in den anderen Kategorien werden von Witschel in den 90er Jahren ausgemacht: als einen der ersten richtungsweisenden rein linguisti-

[9] Eine detaillierte Diskussion dieser Programme findet sich bei Cabré *et al.* 2001.

schen Ansatz bezeichnet er das System *Lexter*, dass 1992 von Bourigault beschrieben wurde, den ersten bedeutenden hybriden Ansatz sieht er bei Daille 1994 (vgl. Witschel 2005:44ff.).

Das erste eigenständige kommerziell vermarktete TEP war laut Fleischmann das Programm *ExtraTerm* der Firma *Trados* und erschien im Jahre 2001 (vgl. 2006:58).

2.3 Charakteristische Merkmale für die Terminologieidentifikation

Die grundsätzliche Aufgabe aller TEP besteht darin, potentielle Termini zu identifizieren, repräsentieren und zur Weiterverarbeitung zur Verfügung zu stellen. Eine wichtige Frage ist für alle TEP: wie können die gewünschten potentiellen Termini identifiziert werden. Hieraus ergeben sich neue Fragen: Wie sehen die gewünschten potentiellen Termini aus? Wodurch unterscheiden sie sich von anderen Bestandteilen des Textes? Was gehört alles zu potentiellen Termini hinzu? Die in dieser Arbeit bereits mehrfach betonten unterschiedlichen Ziele der einzelnen Anwender von TEP führen zu unterschiedlichen Antworten auf diese Fragen.

Da charakteristische Merkmale potentieller Termini auf grundlegende Art die Ausrichtung des Ansatzes und somit den Aufbau eines jeden Terminologieextraktionsprogrammes bestimmen, sollen nun einige grundlegende Eigenschaften von gesuchten Termini betrachtet werden. Witschel beschreibt zwei grundsätzlich unterschiedliche Eigenschaften:

> „[Termini] folgen einer anderen statistischen Verteilung als gemeinsprachliche Wörter und sie werden nach bestimmten Mustern gebildet, die immer wieder – teilweise auch sprachübergreifend – auftauchen und die sich linguistisch auf bestimmte Weise charakterisieren lassen“ (vgl. Witschel 2005:16, Z.15-19).

Die auffällige statistische Verteilung bildet die Grundlage für *statistische Terminologieidentifikationsmethoden* (siehe Kap. 2.4.3). Die besonderen linguistisch charakterisierbaren Muster stellen die Basis für *linguistische Terminologieidentifikationsmethoden* dar (siehe Kap. 2.4.4). Für rein statistische Terminologieextraktionsverfahren werden verschiedene statistische Terminologieidentifikationsmethoden miteinander kombiniert, gleiches gilt für linguistische Verfahren, die aus linguistischen Methoden bestehen. Eine kombinierte Nutzung linguistischer und statistischer Identifikationsmethoden wird als *hybrides Verfahren* bezeichnet. Lieske identifiziert noch eine dritte Eigenschaft: besondere Merkmale von potentiellen Termini wie z.B. Großbuchstaben, die als Grundlage für *Heuristiken* verwendet werden (vgl. 2002:113).

Benennungen (als Teile potentieller Termini) können aus einem oder mehreren Wörtern bestehen. Für diese Unterscheidung werden die Bezeichnungen *Einwortbenennung* bzw. *Mehrwortbenennung* verwendet (vgl. Saß 2004:23). Entscheidend für die Unterteilung ist die räumliche Trennung der Wörter, deshalb zählen Komposita und Abkürzungen zu Einwortbenennungen. Typische Mehrwortbenennungen sind z.B. Nominalphrasen (vgl. Witschel

2005:21-26). Einige Terminologieextraktionsprogramme extrahieren nur Mehrwortbenennungen, v.a. linguistische Verfahren, die nach morphosyntaktischen Strukturen suchen.

Zwei weitere wichtige Eigenschaften potentieller Termini sind *unithood* und *termhood.*[10] Unithood ist der innere Zusammenhalt bzw. die Stabilität einer Benennung: Komposita und einzelne Wörter haben einen sehr starken Unithood, aber auch Mehrwortbenennungen können einen starken Unithood aufweisen. Termhood ist die terminologische Relevanz einer Benennung, eine niedrige terminologische Relevanz weisen Wörter der Gemeinsprache auf, Fachtermini haben einen relativ hohen Termhood. Diese Eigenschaften werden durch unterschiedliche Methoden genutzt, um das Extraktionsergebnis von TEP zu verbessern (vgl. Nakagawa 2001:303f.).

Eine wichtige Eigenschaft der gesuchten Termini ist auch ihre Sprache. Sollen Termini aus zwei unterschiedlichen Sprachen extrahiert werden, kann dies auf unterschiedliche Art und Weise bewältigt werden. Die verschiedenen Ansätze werden in Kapitel 2.5 vorgestellt.

Je nachdem, welche Art von Termini gesucht werden sollen, und welche Ressourcen für diese Suche zur Verfügung stehen, können unterschiedliche Identifikationsmethoden miteinander kombiniert werden, um ein Terminologieidentifikationsverfahren zu erstellen (siehe auch Kapitel 2.6).

2.4 Methoden zur Terminologieidentifikation

In diesem Kapitel sollen Methoden zur Vorverarbeitung sowie statistische, linguistische und Bootstrapping-Methoden zur Terminologieidentifikation beschrieben werden. Diese Methoden werden anhand der bereits vorgestellten charakteristischen Merkmale entwickelt, und im Rahmen eines TEP miteinander kombiniert. Mit der Kombination dieser Methoden und dem Aufbau von TEP beschäftigt sich Kapitel 2.6. Hybride Verfahren verwenden sowohl statistische als auch linguistische Methoden und werden in Kapitel 2.6.4 näher betrachtet.

Eine wichtige interne Unterteilung sowohl statistischer als auch linguistischer Methoden wird durch die Begriffe *intrinsisch* (benennungsbasiert) vs. *extrinsisch* (kontextbasiert) möglich. Intrinsische Methoden betrachten die einzelnen Benennungen und vergleichen sie untereinander auf der Suche nach internen charakteristischen Anzeichen, extrinsische Methoden Betrachten die Benennungen als Ganzes im Kontext des Textes bzw. Korpus (vgl. Zielinski & Ramírez-Safar 2005:2). Um auf die Vor- bzw. Nachteile der einzelnen Verfahren aufmerksam zu machen, enthält jedes Unterkapitel auch einen Überblick über Forschungsmeinungen zu den jeweiligen Ansätzen. Diese beziehen sich oft auf die Effektivität der ein-

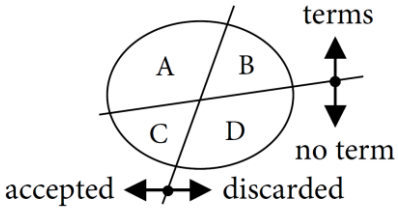
[10] Für Begriffe, für die auch in deutschen Publikationen überwiegend englische Bezeichnungen verwendet werden, und für die sich bisher keine eindeutigen deutschen Entsprechungen durchgesetzt haben, werden auch in dieser Arbeit die englischen Bezeichnungen verwendet.

zelnen Ansätze, insbesondere auf die Qualität des Identifikationsergebnisses, die im nächsten Unterkapitel betrachtet werden soll.

Für diese Arbeit musste zwangsläufig eine Auswahl an Methoden getroffen werden, da aufgrund ihres Umfanges nicht alle Methoden behandelt werden können. Es wird versucht, die grundlegenden Ansätze zu beschreiben, um eine ausreichende Grundlage für die weitere Diskussion der unterschiedlichen Verfahren zu schaffen.

2.4.1 Qualität des Identifikationsergebnisses

Für die quantitative Beschreibung der Qualität des Identifikationsergebnisse von TEP werden vier Begriffe verwendet. Zwei dieser Begriffe werden vor allem für die Bewertung statistischer Verfahren verwendet (*recall* und *precision*) und zwei vor allem für linguistische Verfahren (*noise* und *silence*).



(vgl. Vivaldi & Rodríguez 2007:229)

Für die Berechnung der zugehörigen Werte muss festgestellt werden, welche Termini (A+B) tatsächlich im Korpus enthalten sind (z.B. durch eine manuelle Vergleichsextraktion). Alle potentiellen Termini, die nicht in der Vergleichsextraktion extrahiert wurden, zählen nicht als Termini (C). Die Menge D ist hypothetisch: es handelt sich um nicht vorgeschlagene und auch nicht in der Vergleichsextraktion identifizierte – somit inexistente – Termini. Das Identifikationsergebnis (A+C) besteht aus den verwendbaren Termini (A) und nicht verwendbaren Termini (C). B ist die Menge an Termini, die in der Vergleichsextraktion identifiziert wurden, jedoch nicht im Extraktionsergebnis enthalten waren.

$$recall = \frac{A}{A+B}$$

$$precision = \frac{A}{A+C}$$

$$noise = \frac{C}{A+C}$$

$$silence = \frac{B}{A+B}$$

(vgl. Vivaldi & Rodríguez 2007:229)

Recall beschreibt das Verhältnis gefundener und bestätigter Termini zur Anzahl insgesamt enthaltener Termini. *Precision* beschreibt das Verhältnis bestätigter Termini zur Anzahl der insgesamt gefundenen Termini. *Noise* beschreibt das Verhältnis der verworfenen Termini zur Anzahl insgesamt gefundener Termini. *Silence* beschreibt das Verhältnis nicht gefundener enthaltener Termini zur Anzahl der insgesamt enthaltenen Termini (vgl. Vivaldi & Rodríguez 2006:12). Die Werte können auch prozentual berechnet werden: Prozentsatz des Recalls = A*100/(A+B). Alternativ kann auch nur das Identifikationsergebnis (A+C) betrachtet werden, und die vorgeschlagenen Termini bestätigt bzw. verworfen werden. In diesem Fall können nur Precision bzw. Noise berechnet werden – auf Silence- und Recall-Berechnunen muss verzichtet werden.

Ein weiteres metrisches Maß ist der *F-Wert*, der sich aus Recall und Precision berechnen lässt. Der F-Wert zeigt an, ob sich beide Werte positiv entwickeln, und steigt nur, wenn sowohl Recall als auch Precision groß werden. Das Verhältnis dieser beiden Werte kann je nach gewünschter Gewichtung von Precision und Recall gewählt werden, so kann zum Beispiel der Recall-Wert die doppelte Auswirkung im Vergleich zu jener des Precision-Wertes auf den F-Wert haben (vgl. Witschel 2005:71f.).

2.4.2 Methoden zur Vorverarbeitung

In diesem Kapitel sollen die Methoden zur Vorverarbeitung vorgestellt werden, die zahlreiche Methoden der Korpuslinguistik verwenden. Es handelt sich überwiegend um linguistische Verfahren, die nicht direkt zur Terminologieidentifikation, sondern zur Vorbereitung des Korpus verwendet werden. Diese sollen in den Abschnitten Segmentierung und linguistische Anreicherung betrachtet werden.

2.4.2.1 Segmentierung

Zunächst muss der elektronisch lesbare Text segmentiert werden, da für einen Computer einzelne Sätze oder Wörter sonst nicht erkennbar sind. In einigen Verfahren wird der Text zuerst, z.B. anhand von Satzzeichen in Sätze eingeteilt, dies soll als *Satzsegmentierung* bezeichnet werden.

Ein weiterer Schritt (bzw. in manchen Verfahren der erste Schritt) ist die Unterteilung in einzelne Wörter. Diese Segmentierung wird auch als *Tokenisierung* bezeichnet und nutzt verschiedene Verfahren, um den Text in einzelne *tokens* einzuteilen. Als Grenzen werden üblicherweise Satzzeichen und Leerzeichen verwendet. Teilweise verfügen die Verfahren auch über Abkürzungslisten und ähnliche Hilfsmittel, um z.B. „z.B.“ als einen Token zu erkennen (vgl. Halama 2004:218-224).

Sollen mehrere Token gemeinsam ausgegeben werden, um z.B. nach Mehrwortbenennungen zu suchen, gibt auch die Möglichkeit des *chunkings.* Hierfür werden sogenannte Funktionswörter (auch *non-terms* genannt (vgl. Cabré *et al*. 2001:56)): finite Verben, Pronomina, etc. als Grenzen zwischen sogenannten *chunks* verwendet: Alles was zwischen zwei Funktionswörtern, bzw. zwischen einem Funktionswort und einem Satzzeichen oder zwei Satzzeichen steht, wird als Chunk ausgegeben, und kann in Folge weiter analysiert werden (vgl. Witschel 2005:43f., Halama 2004:228).

Ein segmentierter Text, der aus einer Liste von Sätzen und Token, Token und/oder Chunks besteht, kann bereits für die statistische Weiterverarbeitung verwendet werden. Die Vorverarbeitung für rein statistische Verfahren endet dementsprechend bereits nach diesem Schritt.

2.4.2.2 Linguistische Anreicherung

Für linguistische bzw. hybride Verfahren wird der Text in Folge mit linguistischer Information angereichert. Dies kann durch sogenannte flache Verarbeitungsverfahren (z.B. *POS-Tagging*) oder durch komplexere Verfahren (z.B. *Parsing*) geschehen. Ein Korpus, der bereits mit linguistischer Information angereichert wurde, kann auch als annotierter Korpus bezeichnet werden (vgl. Heid:2004:473).

POS-Tagging (*Part-of-Speech-Tagging* oder auch Wortart-*Tagging*) versucht, jedem Token seine Wortart zuzuordnen. Für Token, auf die mehrere Wortarten zutreffen, wird mittels Disambiguierungsverfahren eine Wortart ausgewählt. Es gibt unterschiedliche Verfahren, die z.B. englischen Texten mit einer Durchschnittsgenauigkeit von über 90% Wortarten zuordnen können (vgl. Halama 2004:224-227). Je komplexer und genauer das Verfahren arbeitet, desto mehr linguistische Ressourcen wie z.B. Lexika bzw. Regelsätze werden benötigt. *POS-Tagger* sind sprachabhängig (vgl. Heid 2006.101f.).

Parsing ist ein komplexes Verfahren zur Erschließung der Syntax und zur Zuordnung von Wortarten in segmentierten Texten. Ein *Parser* versucht jedem Satz seine Syntaxstruktur und somit jedem einzelnen Token seine Funktion innerhalb der Satzhierarchie zuzuordnen (vgl. Langer 2004:252-262). Hierfür müssen grammatikalische Regeln und detaillierte lexikalische Informationen vorhanden sein; denn nur wenn hochspezialisierte Informationen zur Verfügung stehen, kann der Parser ein hochwertiges Ergebnis erzielen (vgl. Heid 2006:104).

Weniger aufwendig als vollständiges Parsing, jedoch komplexer als POS-Tagging funktionieren sogenannte *Chunk-Parser* (bzw. *partielle Parser*, oder *NP-Chunker*). Chunk-Parser parsen Chunks, das heißt, sie analysieren Token-Folgen und ordnen den enthaltenen einzelnen Token syntaktische Funktionen und Wortarten zu (Halama 2004:228-231). Dadurch ist es im Gegensatz zum POS-Tagging, das keine syntaktischen Strukturen analysiert, möglich, festzustellen, ob Token zufällig nebeneinanderstehen oder eine potentielle Mehrwortbenennung sein könnten (vgl. Witschel 2005:43). Darüber hinaus können syntaktische Information und Relationen zwischen einzelnen Benennungen extrahiert werden. Diese Funktionen können auch für die Erstellung von Ontologien verwendet werden (Ahmad & Rogers 2001:750).

Eine weitere Methode der linguistischen Anreicherung ist die *Grundformenbildung* (oder auch Lemmatisierung), die flektierte Wörter in ihre Grundform zurückführt. Dies kann regelbasiert anhand von Heuristiken erfolgen, indem z.B. im Englischen „-ing“ abgeschnitten wird, bzw. „-ies“ durch „y“ ersetzt wird (vgl. Witschel 2005:54f.). Für Sprachen mit einer stärkeren Flexion sind jedoch komplexere Methoden notwendig (vgl. Witschel 2005:100). Eine Möglichkeit sind Vergleichslexika, die beim POS-Tagging konsultiert werden: diese können die entsprechende Grundform enthalten, die dann dem einzelnen Token zusammen

mit seiner Wortart zugeordnet wird (vgl. Evert & Fitschen 2004: 411). Falls keine vollständigen Vergleichslexika zur Verfügung stehen, kann der Vorgang auch halbautomatisch durchgeführt werden (vgl. Heid 2006:102).

2.4.3 Statistische Methoden

Statistische Terminologieidentifikationsmethoden identifizieren potentielle Termini mit einer auffälligen statistischen Verteilung. Die Auffälligkeit dieser statistischen Verteilung kann intrinsisch (Häufigkeit von einzelnen n-Grammen) bzw. extrinsisch (Häufigkeit von einzelnen Ein- bzw. Mehrwortbenennungen) gemessen werden. Darüber hinaus wird unterschieden, ob die Auffälligkeit durch einen Vergleich innerhalb des vorliegenden Textes oder Textkorpus entsteht, oder durch den Vergleich mit einem Vergleichstext bzw. -korpus (vgl. Zielinski & Ramírez-Safar 2005:3).

Grundsätzlich werden die eingelesenen Texte zunächst nach einem bestimmten System, z.B. anhand von Satzzeichen und Leerzeichen, in einzelne Sätze und/oder Token, bzw. längere Chunks und/oder Token geteilt. Dieser Vorgang gehört zur Vorverarbeitung und wird im Kapitel 2.4.2.1 näher behandelt.

Eine sehr simple statistische Methode ist die Bestimmung der *absoluten Frequenz* im betrachteten Text bzw. Textkorpus. Die einzelnen lexikalischen Einheiten werden nach ihrer Frequenz sortiert, eine hohe Frequenz wird als Indiz für eine hohe Relevanz angesehen (vgl. Witschel 2005:17). Durch eine *Mindestfrequenz*, die nur jene Token oder Chunks anzeigt, die eine gewisse Frequenz aufweisen (mind. höher als eins) soll eine höhere Repräsentativität der potentiellen Termini gewährleistet werden und verhindert werden, dass Tippfehler in die Liste potentieller Termini aufgenommen werden (vgl. Witschel 2005:19).

Falls ein Vergleichskorpus verwendet wird, kann die *relative Häufigkeit* von Ein- oder Mehrwortbenennungen im Text bzw. Textkorpus mit der relativen Häufigkeit dieser im Vergleichstext bzw. -korpus verglichen werden. Sollte die relative Häufigkeit höher sein als im Vergleichskorpus gilt dies als Indiz für potentielle Termini (vgl. Heid 2006:102), da Wörter aus der Gemeinsprache üblicherweise keine veränderte relative Häufigkeit verzeichnen. Das Phänomen der erhöhten relativen Häufigkeit wird auch als *Differenzanalyse* (vgl. Witschel 2005:56f.) oder *weirdness* (Ahmad & Rogers 2001:744) bezeichnet. Die Differenzanalyse wird meist extrinsisch angewandt, kann aber auch intrinsisch als Differenzanalyse von n-Grammen, z.B. mit für Fachtermini charakteristischen Trigrammen, verwendet werden (vgl. Witschel 2005:63f.).

Falls kein Vergleichskorpus zur Verfügung steht, funktioniert die relative Häufigkeit ohne weitere Hilfsmittel nicht. Eine Möglichkeit auch ohne Vergleichskorpus zu arbeiten und nicht Unmengen an häufig vorkommenden Wörtern aus der Gemeinsprache zu extrahieren,

sind *Stoppwortlisten*. Stoppwortlisten enthalten Wörter, von denen bereits bekannt ist, dass sie keine terminologische Relevanz haben (vgl. Ahmad & Rogers 2001:741f.). Die entsprechenden Worte im Text bzw. Korpus werden entweder gelöscht oder ignoriert.

Der Unithood kann ebenfalls als statistische Maßzahl als Indiz für Mehrwortbenennungen verwendet werden, indem berechnet wird, wie oft gewisse Token direkt hintereinander im Text enthalten sind. Diese Methode wird auch *mutual information* genannt und vergleicht die Häufigkeit des gemeinsamen Auftretens zweier Token mit der Häufigkeit getrennten Auftretens (vgl. Bowker 2008:64-68). Da die *formale Variation* (also die mögliche unterschiedliche Wortstellung) bzw. die *Trennbarkeit* (Kollokationen müssen nicht unbedingt immer direkt hintereinander vorkommen) in zahlreichen Fällen keine hohen Werte für *mutual information* zulassen, Kollokationen also nicht identifiziert werden können (vgl. Witschel 2005: 26), wird mit der *Kookkurrenz* das gemeinsame Auftreten zweier Token innerhalb eines gewissen Umfelds bemessen (vgl. Ahmad & Rogers 2001:748). Jedoch können eine geringe formale Variation und eine geringe Trennbarkeit auch als gutes Indiz für eine hohe terminologische Relevanz angesehen werden (vgl. Witschel 2005:47).

Auch der Termhood kann statistisch genutzt werden: Eine potentielle Mehrwortbenennung kann einen höheren Wert erhalten, falls einer ihrer Bestandteile bereits als potentielle Einwortbenennung identifiziert wurde (vgl. Witschel 2005:26).

Forschungsmeinungen

Die Häufigkeit ist allgemein als guter Indikator für Termhood anerkannt, jedoch erfordern die relativ hohen Noise-Werte weitere Entwicklungen (vgl. Cabré *et al*. 2001:83). Die Mindestfrequenz kann in kurzen Texten bzw. kleinen Textkorpora oder in Texten, die ein hohes Vorwissen voraussetzen zu hohen Silence-Werten führen, da potentielle Termini, die nur ein- bzw. zweimal vorkommen, von der Extraktion ausgeschlossen werden (vgl. Witschel 2005: 97). Dadurch werden selten vorkommende Termini von statistischen Verfahren nur selten gefunden (vgl. Zielinski & Ramírez-Safar 2005:3). Cabré betont, dass rein statistische TEP sehr große Datenmengen benötigen, um brauchbare Ergebnisse zu liefern (vgl. 2009: 172). Darüber hinaus sind in den Ergebnissen rein statistischer TEP zahlreiche potentielle Termini mehrfach enthalten, in Form von deklinierten Substantiven, gebeugten Verben u.Ä. Das kann nur durch eine linguistische Komponente, die Grundformenbildung verhindert werden (siehe hierzu Kapitel 2.4.2.2).

Als Vorteil der statistischen Ansätze wird angesehen, dass sie keine linguistischen Ressourcen verwenden und somit sprachunabhängig sind (Cabré *et al*. 2001:55). Jedoch gilt dies nicht für statistische TEP, die z.B. Stoppwortlisten verwenden, da es sich dabei – streng genommen – ebenfalls um linguistische Informationen handelt (Cabré *et al*. 2001:79).

2.4.4 Linguistische Methoden

Linguistische Terminologieidentifikationsmethoden[11] identifizieren potentielle Termini, die nach besonderen linguistisch charakterisierbaren Mustern gebildet wurden. Ein wichtiger Bestandteil linguistischer Verfahren ist die Vorverarbeitung, die bereits in Kapitel 2.4.2 vorgestellt wurde. In diesem Kapitel soll auf den sogenannten Musterabgleich, ein linguistisches Terminologieidentifikationsverfahren, das aus verschiedenen Methoden bestehen kann, eingegangen werden.

Musterabgleich

Die linguistischen Identifikationsmethoden können unter dem Oberbegriff Musterabgleich (*pattern matching*) zusammengefasst werden (vgl. Wanhoff 2007:56). Mithilfe dieser Methoden wird innerhalb des linguistisch angereicherten Textes oder Korpus nach bestimmten linguistischen Mustern gesucht. Die Identifizierung dieser linguistischen Muster kann intrinsisch (morphologische Beschaffenheit von Einwortbenennungen bzw. morphosyntaktische Beschaffenheit von Mehrwortbenennungen) oder extrinsisch (Schlüsselwörter im Kontext bzw. morphosyntaktische Beschaffenheit des Kontextes) erfolgen (vgl. Saß 2004:23).

Bei der morphologischen Beschaffenheit von *Einwortbenennungen* wird nach charakteristischen Affixen oder Morphemen gesucht. Diese Methode ähnelt der statistischen Analyse von n-Grammen, benötigt jedoch für ihre Durchführung Informationen über die morphologische Beschaffenheit der zu erfassenden Terminologie und ist im Gegenzug unabhängig von der Frequenz, findet demnach auch potentielle Termini mit einer geringen absoluten Frequenz (vgl. Wanhoff 2007:56ff.). Die Methode eignet sich für die Extraktion von Termini aus Fachbereichen mit besonders hohem Anteil an charakteristischen Wortbestandteilen (vgl. Witschel 2005:83). Es kann auch generell nach Komposita gesucht werden (vgl. Zielinski & Ramírez-Safar 2005:3).

Eine oft angewandte Methode ist die Suche nach *Mehrwortbenennungen* mithilfe morphosyntaktischer Muster. Für diese Methode werden Muster häufig vorkommender Mehrwortbenennungen gebildet (z.B. Adjektiv+Nomen, Nomen+Präposition+Nomen, etc.). Die Bezeichnungen der einzelnen Bestandteile müssen den bei der linguistischen Anreicherung vergebenen Wortartbezeichnungen entsprechen; dann können diejenigen potentiellen Mehrwortbenennungen herausgefiltert werden, die den Mustern entsprechen (vgl. Wanhoff 2007:57). Falls bei der linguistischen Anreicherung ein Parser verwendet wurde, kann verhindert werden, zufällig nebeneinanderstehende, dem Muster entsprechende Token als potentielle Mehrwortbenennung zu identifizieren (vgl. Witschel 2005: 43).

[11] Die linguistischen Methoden werden auch als morphologische, musterbasierte bzw. computerlinguistische Methoden bezeichnet. Im Rahmen dieser Arbeit wird die Bezeichnung linguistisch bevorzugt.

Die *extrinsischen Methoden* des Musterabgleichs werden relativ selten verwendet. Saß hat zwei Methoden identifiziert: Die Suche nach Schlüsselwörtern und die Analyse der morphosyntaktischen Beschaffenheit. Schlüsselwörter im Kontext sind z.B. „*als X bezeichnet man* oder *ein sogenanntes Y*“ (vgl. Saß 2004:24, Z. 10f.). X und Y stellen potentielle Termini dar. Auch Klammerinhalte bzw. Bestandteile von Aufzählungen werden als potentielle Termini angesehen. Für dieses Verfahren sind jedoch aufwendige Vorarbeiten notwendig (vgl. Saß 2004:24). Methoden zur Analyse der morphosyntaktischen Beschaffenheit des Kontextes sind jene, die alle Chunks zwischen Funktionswörtern extrahieren bzw. die Zwischenergebnisse mit bereits bestätigten Termini aus Terminologiedatenbanken abgleichen (vgl. Saß 2004:24).

Forschungsmeinungen

Als allgemeiner Nachteil linguistischer Methoden wird die starke Sprachgebundenheit angesehen (vgl. Bowker 2008:84). Die für die linguistische Anreicherung notwendigen Informationen sind bisher nur in wenigen Sprachen (z.B. Englisch) vorhanden (vgl. Zielinski & Ramírez-Safar 2005:3).

Cabré *et al*. kritisieren bei Mehrwortbenennungen, die per morphosyntaktischen Musterabgleich gesucht wurden, den hohen Noise-Anteil durch zufällig nebeneinanderstehende Token, die dem Muster entsprechen, sowie die hohen Silence-Werte, die durch starre Suchmuster verursacht werden (vgl. Cabré *et al.* 2001:83). Eine Verbesserung dieser Werte wäre durch eine syntaktische Analyse möglich, das vollständige Parsen ist jedoch sehr aufwendig und somit auch teuer (Heid 2006:104). Um trotzdem die Vorteile einer syntaktischen Analyse gegenüber einer einfachen POS-Filterung nutzen zu können, werden zahlreiche Zwischenlösungen entwickelt, wie unter Chunk-Parsing (siehe Kap. 2.4.2) beschrieben (vgl. Witschel 2005:43-46, auch Heid 2006:106).

Zielinski & Ramírez-Safar attestieren linguistischen Methoden im Vergleich zu statistischen eine bessere Unterteilung der Token in potentielle Termini und befürworten die Grundformenbildung, die zahlreiche Wiederholungen im Identifikationsergebnis verhindert (vgl. 2005:3).

2.4.5 Bootstrapping-Methoden

Um die Identifikation der potentiellen Termini effizienter zu gestalten, wurden einige Ansätze entwickelt, die nicht eindeutig linguistisch oder statistisch klassifizierbar sind. Eine Gemeinsamkeit dieser Ansätze ist die Einteilung der Identifikation in mehrere Schritte. Im ersten Schritt werden durch eine Kombination der oben vorgestellten Methoden potentielle Benennungen extrahiert. Diese werden dann in weiterer Folge genutzt, um erneut potentielle

Benennungen zu extrahieren. Zwischen diesen beiden Schritten kann der Anwender bei einigen Methoden eingreifen (z.B. *relevance feedback*), bei anderen Methoden verlaufen die Schritte intern und der Anwender bekommt erst das Endergebnis zu sehen. Die Verwendung einer sogenannten Startmenge, die in mehreren Identifikationsvorgängen ausgeweitet wird, wird auch als *bootstrapping* bezeichnet (vgl. Witschel 2005:50).

Für eine Anwendung des Relevance Feedbacks wird dem Anwender nach einem ersten Identifikationsdurchgang eine Liste potentieller Benennungen präsentiert, aus welcher der Anwender die für ihn relevanten Benennungen auswählen kann. Im zweiten Durchgang werden dann den ausgewählten Benennungen ähnliche Benennungen höher gewertet und ebenfalls extrahiert und den verworfenen Benennungen ähnliche Benennungen niedriger gewertet und somit verworfen. Wie diese erneute Suche und Umsetzung aussieht, hängt von den einzelnen verwendeten Methoden ab. Es können zum Beispiel im ersten Schritt Einwortbenennungen und im zweiten Schritt Mehrwortbenennungen extrahiert werden. Recall- sowie Precision-Werte können durch diese Methode erhöht werden, da die Ergebnisse an die Anwenderbedürfnisse angepasst werden können (vgl. Witschel 2005:32f.).

Einen ähnlichen Ansatz verfolgt Heid mit der Identifikation von *domänenspezifischen Morphemen*. Im Rahmen einer im ersten Schritt vorgenommenen Differenzanalyse werden die häufigsten terminologisch relevanten Einwortbenennungen eines Textes bzw. Korpus identifiziert. Die Komposita unter diesen Benennungen werden in ihre Morpheme zerlegt. Alle simplen Einwortbenennungen und Morpheme, die eine gewisse Mindestfrequenz aufweisen, werden als domänenspezifisch erachtet. Bei der weiteren Identifikation werden alle Komposita bzw. potentiellen Mehrwortbenennungen höher gewertet, die einen *domänenspezifisches Morphem* enthalten (vgl. Heid 1998). In diesen Prozess wird der Anwender nicht mit einbezogen, er muss dementsprechend keine Zwischenergebnisse validieren.

Falls bereits validierte Terminologie vorhanden ist, kann diese ebenfalls verwendet werden, um das Identifikationsergebnis zu verbessern, und neue Benennungen bzw. Varianten bekannter Termini zu finden (vgl. Cabré *et al*. 2001:56).

2.5 Methoden zur computergestützten bilingualen Terminologieextraktion

Mithilfe von Terminologieextraktionsprogrammen können sowohl monolinguale als auch bilinguale Terminologien erstellt werden. Für eine präzise Differenzierung ist es wichtig zwischen dem Extraktionsvorgang selbst und dem angestrebten Endresultat zu unterscheiden: Die Erstellung einer bilingualen Terminologie erfordert nicht unbedingt eine bilinguale Terminologieextraktion, die Termini der zweiten Sprache können auch nachträglich im Zuge der Weiterverarbeitung zu einem Extraktionsergebnis einer monolingualen TE hinzugefügt werden.

Bei der computergestützten *monolingualen Terminologieextraktion* werden einsprachige Texte bzw. Textkorpora verwendet, um darin einsprachige potentielle Termini zu identifizieren. Das Ziel einer monolingualen TE kann sowohl die Erstellung bzw. Erweiterung eine monolingualen als auch einer bilinguale Terminologie sein. Im Gegensatz zur bilingualen Terminologieextraktion gibt es für die monolinguale keine besonderen Methoden, sondern es können alle in Kapitel 2.4 vorgestellten Methoden verwendet werden.

Bei der computergestützten *bilingualen Terminologieextraktion* werden Texte bzw. Textkorpora in zwei verschiedenen Sprachen verwendet, um darin potentielle Termini zu identifizieren. Ziel einer zweisprachigen Terminologieextraktion ist in der Regel die Erstellung einer zweisprachigen Terminologie. Für die bilinguale TE werden zusätzliche Methoden benötigt, die in Kapitel 2.4 vorgestellten Methoden bieten jedoch oft die Grundlage für einen bedeutenden Teil von bilingualen Terminologieextraktionsprogrammen.

Computergestützte Extraktion von Terminologiepaaren

Es gibt unterschiedliche Methoden für die Behandlung zweisprachiger Ressourcen. Als Grundlage benötigen fast alle Methoden *parallele* Texte bzw. Textkorpora, die jeweils in zwei Sprachen vorliegen, also Übersetzungen voneinander sind (vgl. Thurmair 2003:173, Witschel 2005:119f.). Die Texte sind bzw. werden segment-, satz- oder absatzweise *aligniert*, das bedeutet, dass jeweils ein Element in einer Sprache dem entsprechenden Element in der anderen Sprache zugeordnet ist bzw. wird. Alignierte Texte finden sich z.B. in *translation memories,* es besteht aber auch die Möglichkeit, Texte für eine computergestützte Terminologieextraktion zu alignieren.

Bilinguale Terminologieextraktionsprogramme versuchen, innerhalb der alignierten Elemente Terminologiepaare (zusammengehörige bzw. sich entsprechende Termini) zu identifizieren. Dies kann gleichzeitig geschehen oder aber gestaffelt, indem erst eine monolinguale Terminologieextraktion im einem Text durchgeführt wird, und im Anschluss entsprechende potentielle Entsprechungen im zweiten Text gesucht werden (vgl. Thurmair 2003: 170). Heid beschreibt eine weitere Methode, in der zunächst monolinguale Terminologieextraktionen in beiden Sprachen durchgeführt werden und anschließend versucht wird, durch *alignment* die entsprechenden Paare zu finden (vgl. 2004:478). In der von Thurmair beschriebenen Methode der sprachübergreifenden Identifikation (*crosslingual retrieval*) wird mit nicht-alignierten Texten gearbeitet, die dieselben Themenbereiche abdecken (vgl. 2003:173).

Forschungsmeinungen

Vor allem im Rahmen einer begriffsorientierten Terminologiearbeit bietet sich auch für die Erarbeitung bilingualer Terminologien eine monolinguale Extraktion an bzw. wird zwingend vorausgesetzt (vgl. Fleischmann 2006:13).

Zielinski & Ramírez-Safar identifizieren im Übersetzungsprozess unterschiedliche Anwendungsmöglichkeiten und Zeitpunkte für die bilinguale bzw. monolinguale TE: vor der Übersetzung wird durch die monolinguale TE eine Terminologieliste für die Vorbereitung der Übersetzung erstellt und die bilinguale TE wird genutzt, um Terminologie aus bereits übersetzen Texten zur Wiederverwendung zu extrahieren (vgl. 2005:2).

Bowker stellt fest, dass die meisten TEP für monolinguale TE konzipiert sind (vgl. 2008:82). Forschungsarbeiten reflektieren dieses Phänomen, indem sie oft hauptsächlich monolinguale TE betrachten und nur am Rande auf bilingualer TE eingehen (vgl. z.B. Witschel 2005). Ein Grund hierfür könnte auch sein, dass zahlreiche bilinguale TEP fast vollständig auf monolingualen Terminologieextraktionsverfahren basieren, und nur darüber hinaus ergänzende Funktionen haben, um die Vervollständigung der zweiten Sprache zu erleichtern; und neue Erkenntnisse für die monolinguale TE sich dementsprechend auch für die bilinguale TE anwenden lassen.

Fleischmann betont, dass die Ergebnisse monolingualer TE in der Regel besser sind, als die Ergebnisse bilingualer TE (vgl. 2006:55). Außerdem kritisiert er, dass ein Teil der Terminologiepaare aus Übersetzungen entnommen wird, die keine autoritativen Quellen darstellen Fleischmann (vgl. 2006:13).

2.6 Aufbau von Terminologieextraktionsprogrammen

Alle Terminologieextraktionsprogramme identifizieren potentielle Benennungen aus einem elektronisch lesbaren Text bzw. Korpus und präsentieren diese (evtl. zusammen mit weiteren extrahierten Informationen) zur Weiterverarbeitung dem Anwender (vgl. Cabré *et al.* 2001: 54). Sie führen nacheinander mehrere Aufgaben aus:

(1) Einstellungen abfragen,
(2) einen Text bzw. Textkorpus einlesen,
(3) segmentieren, tokenisieren bzw. chunken,
    (a) (evtl. linguistisch anreichern: pos-taggen oder parsen, lemmatisieren),
(4) potentielle Benennungen identifizieren, reihen, filtern (statistisch, linguistisch oder hybrid),*
(5) potentielle Benennungen anzeigen,
    (a) (evtl. validieren lassen, erneut mit (4) beginnen),
    (b) (evtl. zusätzliche Information extrahieren und ebenfalls anzeigen),

(6) Möglichkeiten zur Weiterverarbeitung
(7) und zum Export bieten.

*Falls Bootstrapping-Methoden angewendet werden, wiederholt sich Punkt (4).

Unterschiedliche Komponenten übernehmen die einzelnen Aufgaben. Es können vier Komponenten identifiziert werden, die für die genannten Aufgaben zuständig sind:

1. Interaktion mit dem Anwender (1)+(5+a+b)+(6)+(7),
2. Dateneingabe und -ausgabe (2) +(7),
3. Vorverarbeitung (3+a),
4. Identifikation (4).

Die einzelnen Komponenten sollen in Folge einzeln betrachtet werden. Die Beschreibung der Komponenten soll hier nicht der Modellierung einer Software dienen, sondern lediglich veranschaulichen, welche verschiedenen Teile ein TEP gewöhnlich enthält, und welche unterschiedlichen Möglichkeiten sich durch die Kombination verschiedener Komponenten ergeben.

2.6.1 Interaktion mit dem Anwender

Die *Interaktion mit dem Anwender* kann unterschiedlich gestaltet werden und ist stark abhängig von den Möglichkeiten eines TEP, die einzelnen Komponenten zu individualisieren und somit das Programm für bestimmte Zwecke anzupassen. Eine erste Möglichkeit zur Interaktion bietet sich dem Anwender durch die Parametereinstellung noch vor der Identifikation potentieller Benennungen. Während der Identifikation kann der Anwender nur selten Einfluss auf den Verlauf nehmen, diese Möglichkeit wird ihm lediglich durch Relevance Feedback gewährt. Nach der Identifikation findet erneut eine Interaktion mit dem Anwender statt, dazu gehören die Repräsentation, Weiterverarbeitung und der Export der Ergebnisse.

2.6.1.1 Parametereinstellung

Im Zuge der Parametereinstellung können gewisse Parameter für die TE eingestellt bzw. Quellen für linguistische Daten angegeben werden. Einstellbare Parameter sind z.B. die minimale bzw. maximale Länge von potentiellen Benennungen, die maximale Anzahl an Token innerhalb einer Benennung, die Mindestfrequenz, die maximale Anzahl zu extrahierender Benennungen, die zu verwendenden Muster, die Beachtung bzw. Nichtbeachtung von Wortarten, bzw. bei bilingualer Extraktion die Anzahl der angebotenen Übersetzungen (vgl. Zielinski & Ramírez-Safar 2005:5, Saß 2004:70). Falls das Programm über Filter für die Verarbeitung des Textes (siehe Kapitel 2.6.2) verfügt, können diese eventuell ausgewählt werden.

Die Parametereinstellung kann das Identifikationsergebnis entscheidend beeinflussen. Dies soll am Beispiel der Mindestfrequenz erklärt werden: Falls im Text besonders viele wenig frequente Benennungen vorkommen, muss die Mindestfrequenz dementsprechend niedrig sein; sollen jedoch nur repräsentative Benennungen extrahiert werden, muss dieser Wert hoch eingestellt werden. Auch längere Texte bzw. große Korpora verlangen eine höhere Mindestfrequenz (vgl. Witschel 2005: 112f.).

2.6.1.2 Relevance Feedback

Falls die Methode des Relevance Feedback angewendet wird (siehe auch Kapitel 2.4.5), bekommt der Anwender ein Zwischenergebnis der Identifikation präsentiert, das von ihm validiert werden muss. Anhand der getroffenen Entscheidungen wird die Identifikation fortgesetzt. Durch diesen Vorgang kann das Ergebnis grundlegender an die Benutzerwünsche angepasst werden und in der Qualität verbessert werden (vgl. Witschel 2005: 32f.).

2.6.1.3 Repräsentation und Weiterverarbeitung der Ergebnisse

Das Ergebnis der Identifikation (siehe Kapitel 2.6.4) wird dem Anwender in Form der Repräsentation angezeigt. Dies kann in Form einer einfachen Liste geschehen, aber auch als Datenbank mit Abfragemöglichkeiten. Die Ergebnisse können je nach Möglichkeiten nach verschiedenen Kriterien geordnet (z.B. Häufigkeit) oder sortiert (z.B. Einwort- getrennt von Mehrwortbenennungen bzw. nach Alphabet) werden, manche Programme ermöglichen auch eine Suche innerhalb der Ergebnisse (vgl. Zerfaß 2006:22-24).

Zusätzliche Informationen können z.B. durch eine sogenannte KWIC-Suche (*Keyword-in-context-Suche*) bereitgestellt werden. Hierbei wird im Originaltext bzw. -korpus nach der potentiellen Benennung gesucht und die entsprechenden Textstellen werden dem Anwender angezeigt (vgl. Zielinski & Ramírez-Safar 2005:4). Diese Funktion ermöglicht es in manchen Fällen auch, potentielle Termini mit zugehörigen Originalkontexten zu extrahieren (vgl. Zerfaß 2006:22f.).

Darüber hinaus können Beziehungen zwischen potentiellen Termini angezeigt werden, falls die Identifikationskomponente Möglichkeiten dazu bereitstellt. Dies ist zum Beispiel bei vollständig oder partiell geparstem Text möglich. Diese Informationen können beispielsweise für die Erstellung von Ontologien verwendet werden (vgl. Ahmad & Rogers 2001:750).

Die Möglichkeiten der Weiterverarbeitung sind sehr unterschiedlich. In integrierten Programmen können teilweise nur Operationen vorgenommen werden, die dem nächsten Workflowschritt entsprechen, z.B. die potentiellen Termini sofort validieren, übersetzen und mit der Übersetzung fortsetzen. Da die jeweiligen Zusatzinformationen (z.B. Kontext) nicht immer gemeinsam mit den potentiellen Termini extrahiert werden können, zwingen einige

Programme ihre Anwender, die Termini bereits vor dem Export zu validieren, bzw. bereits vor dem Export Definitionen zu verfassen (Zerfaß 2006:22-24). Dies entspricht jedoch nicht immer der Arbeitsteilung in der Terminologieerfassung, in der z.B. die Extraktion von Terminologen durchgeführt wird und Experten zur Validierung hinzugezogen werden (vgl. Fleischmann 2006:64ff).

Die bilinguale Extraktion kann auch über die Interaktion mit dem Anwender durchgeführt werden. Die KWIC-Suche kann dem Anwender Kontexte anzeigen, auf Grundlage derer er sich für eine vorgeschlagene Übersetzung entscheiden kann (vgl. Zerfaß 2006:23).

2.6.2 Dateneingabe und -ausgabe

Die Dateneingabe ist für die erste Verarbeitung des Textes und die Datenausgabe für den Export des Extraktionsergebnisses wesentlich. Hierbei spielen kompatible Dateiformate eine bedeutende Rolle.

Es kann zwischen eigenständigen Programmen und in eine größere Umgebung (z.B. in ein CAT-Programm) integrierten Programmen unterschieden werden. Bei integrierten Programmen muss evtl. ein Übersetzungsprojekt angelegt werden, in einem weiteren Schritt steht der Text dann für eine Extraktion zur Verfügung. In diesem Fall hängt die Kompatibilität vom CAT-Programm ab (vgl. Zerfaß 2006:22f.). Bei eigenständigen Programmen ist ebenfalls festgelegt, welche Dateiformate gelesen werden können. Diese Möglichkeiten können stark divergieren und eine eventuelle Konvertierung erfordern. Ein weiterer Aspekt ist der Umgang mit Tags, z.B. HTML-Tags, die, wenn sie nicht gefiltert werden, zu irrelevanten Zeichenfolgen im Extraktionsergebnis führen können, und somit die Qualität des Identifikationsergebnisses negativ beeinflussen oder die Weiterverarbeitung erschweren können (vgl. Gasser 2004: 22f.).

Die Exportmöglichkeiten von TEP können ebenfalls stark variieren, TEP die nicht eigenständig sind, sondern z.B. in ein CAT-Tool integriert sind, bieten kaum Exportmöglichkeiten an (vgl. Zerfaß 2006:23). Eigenständige TEP bieten oft einen direkten Export in Datenbanken oder aber in gängige Dateiformate wie z.B. CSV (Trennzeichenformat) oder XLS (Microsoft Office Excel) an (vgl. Zerfaß 2006:22-24).

2.6.3 Vorverarbeitung

Für die Vorverarbeitung werden die in Kapitel 2.4.2 vorgestellten Methoden verwendet: entweder wird der Text tokenisiert und für eine linguistische Verarbeitung werden in Folge die Wortarten annotiert, oder er wird gechunkt und für eine linguistische Verarbeitung partiell geparst; vollständiges Parsen ist eine Seltenheit. Auch für rein statistische Verfahren muss der Text tokenisiert oder gechunkt werden. Je nach Verfügbarkeit wird für diesen Vorgang

auf Stoppwort- oder Funktionswort-Listen, Lexika, Vergleichskorpora und weitere linguistische Informationen zurückgegriffen. Die Qualität der Vorverarbeitung wirkt sich direkt auf das Identifikationsergebnis aus (vgl. Heid 2004: 474).

2.6.4 Identifikationsverfahren

Die Identifikation setzt sich aus mehreren Methoden zur Terminologieidentifikation zusammen, viele Verfahren sind laut Heid zweistufig (vgl. 2006:103f.). Natürlich kann theoretisch auch nur eine einzige Methode verwendet werden, dieser Fall soll hier jedoch nicht näher betrachtet werden.

Für ein zweistufiges Verfahren werden erst eine oder mehrere Methoden verwendet, um eine erste Liste an potentiellen Benennungen zu erstellen, die meist auch schon anhand ihrer vermutlichen Relevanz in eine Reihenfolge (*ranking*) gebracht werden (*erste Identifikation*). Aus dieser Liste wird dann durch eine oder mehrere Methoden eine erneute Auswahl getroffen (*Filterung*) - manchmal erfolgt das Ranking auch erst an dieser Stelle. (Hier kann nun evtl. ein zweiter (oder x-ter) Identifikationsschritt einsetzen.) Schließlich werden alle oder eine bestimmte Anzahl des Ergebnisses (*Auswahl*) als potentielle Termini ausgeben (*Repräsentation*). Die Repräsentation ist bereits Teil der Interaktion mit dem Benutzer (siehe Kapitel 2.6.1). (Im Fall von Relevance Feedback beginnt die *Identifikation* nach einer Interaktion mit dem Benutzer erneut.)

Bootstrapping-Methoden verwenden den ersten Identifikationsschritt für die Erstellung einer Startmenge und den zweiten Identifikationsschritt für eine erste Ausweitung dieser Startmenge. Es können auch noch weitere Identifikationsschritte eingesetzt werden, um die neue Startmenge nochmals zu erweitern (vgl. Witschel 2005:65).

Bilinguale Terminologieextraktionsprogramme setzen entweder direkt in der ersten Identifikation besondere Methoden ein, oder aber nutzen den zweiten Identifikationsschritt für die Identifikation der potentiellen Terminologiepaare (vgl. Thurmair 2003:173).

Die in Kapitel 2.4 vorgestellten Identifikationsmethoden können zu rein linguistischen, rein statistischen oder hybriden Verfahren kombiniert werden. Ahmad & Rogers stellen fest, dass die meisten TEP hybride Verfahren verwenden (vgl. 2001:746), unter kommerziellen Produkten finden sich neben vielen hybriden auch zahlreiche statistische Programme (Zielinski & Ramírez-Safar 2005:5).

Hier zur Veranschaulichung ein fiktives Beispiel[12] für ein hybrides Programm mit zweifacher Identifikation (inkl. Verwendung von Bootstrapping, Rankings und Filterung): ein Text wird tokenisiert und pos-getagged (*Vorverarbeitung*), dann wird mithilfe eines Ver-

[12] Dieses Beispiel entspricht keinem der Autorin bekannten TEP, deckt sich jedoch teilweise (in vereinfachter Form) mit dem von Witschel entwickelten TEP (vgl. Witschel 2005).

gleichskorpus eine Differentialanalyse für Einwortbenennungen durchgeführt (*erste Identifikation mit Ranking*), die ersten 100 werden als Startmenge festgelegt (*Auswahl*). In Folge werden anhand von morphosyntaktischen Mustern Mehrwortbenennungen gesucht (*zweite Identifikation*), diese werden nach Häufigkeit und Beinhaltung einer ausgewählten Einwortbenennung gerankt (*Filterung mit Ranking, Verwendung von Bootstrapping*) und die ersten 100 der Liste (*Auswahl*) werden zusammen mit der Auswahl der Einwortbenennungen als Ergebnis ausgegeben (*Repräsentation*).

2.7 Zusammenfassung

Im Rahmen dieses Kapitels wurden Terminologieextraktionsprogramme betrachtet. Zu Beginn des Kapitels wurden die Gründe für die Entwicklung von TEP erläutert und ein Einblick in die Entwicklungsgeschichte gegeben. Darauffolgend wurden die charakteristischen Merkmale potentieller Termini vorgestellt, die für die Terminologieidentifikation genutzt werden. Methoden zur Vorverarbeitung sowie statistische, linguistische und Bootstrapping-Methoden zur Terminologieidentifikation wurden erklärt und ihre Vor- und Nachteile aufgezeigt. Anschließend wurden unterschiedliche Ansätze zur bilingualen TE betrachtet und zum Abschluss der Aufbau von Terminologieextraktionsprogrammen anhand der einzelnen zu bewältigenden Aufgaben und Komponenten dargestellt.

Das nächste Kapitel widmet sich der computergestützten Terminologieextraktion in der Praxis unter besonderer Berücksichtigung bekannter Probleme.

3 Computergestützte Terminologieextraktion in der Praxis

In diesem Kapitel sollen die wesentlichen praktischen Aspekte der computergestützten TE betrachtet werden. Zunächst werden die einzelnen Phasen der computergestützten Terminologieextraktion betrachtet (3.1), anschließend wird auf weitere Anwendungsmöglichkeiten eingegangen (3.2). Kapitel 3.3 beschäftigt sich mit Problemen bei der Anwendung von TEP und Kapitel 3.4 widmet sich Desiderata für die weitere Entwicklung von Terminologieextraktionsprogrammen.

3.1 Anwendung von Terminologieextraktionsprogrammen

Die Anzahl und das Ausmaß der einzelnen Arbeitsschritte kann je nach TEP und Art der Anwendung des TEP stark variieren, jedoch können alle Arten der Anwendung in drei Phasen eingeteilt werden: Vorbereitung, Identifikation, Nachbearbeitung (vgl. Lieske 2002:111). In diesen drei Phasen werden zahlreiche Arbeitsschritte vollzogen, u.a. die Korpuserstellung, Parametereinstellung, Validierung und Bearbeitung der potentiellen Termini sowie der Export und die Weiterverarbeitung (vgl. Ahmad & Rogers 2001:740, Lieske 2002:111f.). Unterteilt in die drei Phasen soll in Folge ein Überblick über diese einzelnen Schritte gegeben werden. Nicht jeder Arbeitsschritt wird bei jeder Art von Anwendung ausgeführt, dennoch sollen die verschiedensten Möglichkeiten präsentiert werden, um einen möglichst umfassenden Überblick zu gewährleisten.

3.1.1 Vorbereitungsphase

Vor allem in der Vorbereitungsphase sind die unterschiedlichen Ausgangssituationen und Ziele der jeweiligen TE von großer Bedeutung. Eine einmalige Terminologieextraktion unterscheidet sich von einer wiederholten bzw. regelmäßigen TE in der Planung, im Ausmaß der Vorarbeit und auch in den bereits vorhandenen Ressourcen (vgl. Fleischmann 2006:48, Wohlfahrt 2008:7). In manchen Fällen stehen auch bei einer einmaligen oder erstmaligen (später zu wiederholenden) TE bereits befüllte Terminologiedatenbanken oder andere verwendbare Ressourcen zur Verfügung, in anderen Fällen wurde noch keinerlei Vorarbeit erledigt. Hierbei ist zu beachten, dass nicht alle TEP durch Terminologieerkennungsmethoden die Inhalte bereits vorhandener Terminologiedatenbanken berücksichtigen können (vgl. Zielinski & Ramírez-Safar 2005:24).

Auch eine bereits zum wiederholten Male regelmäßig durchgeführte TE erfordert eine bestimmte Vorbereitung, die sich jedoch von einer erstmaligen unterscheidet: gewisse Entscheidungen wurden bereits getroffen, das Augenmerk wird bei der Erfassung neuer Termini

vor allem auf einer möglichst effizienten Nutzung und Überprüfung der bereits vorhandenen Ressourcen liegen (vgl. Reuther & Rösener 2007:44).

3.1.1.1 Planung der Extraktion

In einem ersten Schritt sollten die Ziele (z.B. Erstellung einer firmenspezifischen Terminologie, Erleichterung der Übersetzung, Erstellung eines allgemeinen Glossars) und die den Zielen entsprechenden Charakteristika der Extraktion (z.B. punktuell oder systematisch bzw. einfach, wiederholt oder regelmäßig) festgelegt werden (vgl. Gasser 2004:5, Zerfaß 2008: 12f.). Gemäß den einzelnen Charakteristika müssen weitere Entscheidungen getroffen werden (z.B. das Intervall der regelmäßigen TE, wer welche Aufgaben durchführt, etc.), in diesem Rahmen kann auch ein vollständiger Workflow mit bereits festgelegten Terminen entwickelt werden (vgl. König 2009:47). Das bedeutet natürlich nicht, dass dies bei jeder Terminologieextraktion auch wirklich strategisch überdacht wird – in der Praxis fällt die theoretische Vorbereitung sicherlich unterschiedlich detailliert aus.

3.1.1.2 Korpusbildung

Bei einer *systematischen TE* gilt es, den terminologisch relevanten Bereich zu definieren. Sobald der terminologisch relevante Bereich feststeht, müssen die zu verwendenden Dokumente identifiziert und gesammelt werden, um einen Korpus zu bilden (vgl. Fleischmann 2002:48). Bei einer *punktuellen Terminologieextraktion* ergibt sich der Korpus oft automatisch aus den vorhandenen Dokumenten, die übersetzt werden müssen oder bereits übersetzt wurden (vgl. Zerfaß 2008:10). In diesem Fall wird je nach Zweck der TE kein besonderer Bereich festgelegt, da für übersetzerische Zwecke oft alle problematischen Wörter und Wortgruppen als Termini akzeptiert werden (vgl. Thurmair 2003:171). Für die Erstellung einer themenbezogenen Terminologie kann jedoch auch bei einer punktuellen TE ein terminologischer Bereich definiert werden, bzw. schrittweise anhand von einzelnen Projekten eine umfassende Terminologie erarbeitet werden (vgl. Wohlfahrt 2008:7).

Die Korpusbildung kann unterschiedlich aufwendig ausfallen[13], und je nach Zweck der TE und Identifikationsmethode des TEP müssen unterschiedliche Faktoren beachtet werden: Sollen möglichst viele *ähnliche Texte* verwendet werden, um besonders gute statistische Ergebnisse zu erzielen oder sollen ähnliche Texte vermieden werden, um eine Überbewertung einzelner Bereiche zu verhindern (vgl. Fleischmann 2006:51f.). In diesem Zusammenhang kann auch bereits entschieden werden, ob der Korpus als Ganzes oder in Teilen bearbeitet

[13] Für eine Checkliste eventuell relevanter Dokumente für die Korpusbildung in einem Unternehmen vgl. Fleischmann 2006:49.

werden soll – dies kann ebenfalls Auswirkungen auf das Ergebnis haben (vgl. Fleischmann 2006:51f.).

3.1.1.3 Korpusvorverarbeitung

Auch das *Dateiformat* der Dokumente ist von Bedeutung. Je nach TEP müssen einige Dokumente vor der Verarbeitung umgewandelt werden, einige können gar nicht verarbeitet werden. Für die Verarbeitung von PDF-Dokumenten reicht z.B. das einfache Kopieren des Textes nicht aus, da die besondere Absatzstruktur in PDF-Dokumenten die Satzstrukturen des Textes verzerrt (vgl. Fleischmann 2006:49ff.).

Für den Fall, dass ein TEP für die Vorbereitung bzw. den Extraktionsvorgang *weitere linguistische Ressourcen*, wie z.B. Vergleichskorpora oder Stoppwortlisten benötigt, so müssen diese ausgewählt, erworben oder erstellt werden (vgl. Gasser 2004:15).

3.1.1.4 Auswahl des TEP und Konfiguration

Falls noch kein Terminologieextraktionsprogramm zur Verfügung steht, muss ein entsprechendes Programm erworben oder entwickelt werden. Dies kann je nach Anforderungen unterschiedlich kosten- bzw. zeitintensiv sein, und sollte erst im Anschluss an eine Analyse der Ausgangssituation und der Ziele der TE geschehen (vgl. Fleischmann 2006:45). Sobald das Programm vorhanden ist, kann nach einer Einarbeitung mit der Einstellung der Parameter des Extraktionsprogrammes begonnen werden. Je nach Programm bieten sich wenige bis zahlreiche komplexe Einstellmöglichkeiten[14], die zum Teil starke Auswirkungen auf das Extraktionsergebnis haben können (vgl. Saß 2004:79-85). Falls hierfür keine Erfahrungswerte bzw. Vorgaben vorliegen, müssen unter Umständen Testextraktionen zur Ermittlung geeigneter Einstellungen durchgeführt werden (vgl. Gasser 2004:29-32).

3.1.2 Identifikationsphase

Sobald die eventuell notwendigen Testdurchläufe beendet und alle Parameter entsprechend eingestellt wurden, beginnt die Identifikationsphase. In dieser Phase werden die Dokumente eingelesen und durchlaufen die einzelnen Stufen des verwendeten Programms (Vorverarbeitung, Identifikation, Darstellung, vgl. Kapitel 2.6). Je nach Korpus und Extraktionsanwendung kann die Identifikation selbst in einem Durchgang (gesamter Korpus) oder in mehreren Etappen (einzelne Texte bzw. Teile des Korpus) durchgeführt werden. Dies hängt eventuell auch davon ab, ob es für den Anwender Eingriffsmöglichkeiten in den Identifikationsprozess gibt: Wie in Kapitel 2.4.5 beschrieben, müssen bei einigen Bootstrapping-Identifikationsverfahren während der Identifikation Zwischenergebnisse validiert werden.

[14] Für eine Übersicht der Parameter von kommerziellen TEP vgl. Zielinski & Ramírez-Safar 2005:5f.

3.1.2.1 Beteiligte Akteure

Falls mehrere Akteure am Extraktionsprozess beteiligt sind, ist es wichtig, dass die Vorgaben für alle Parameter und Entscheidungen im Vorhinein schriftlich fixiert werden und von allen Beteiligten eingehalten werden, da modifizierte Parameter das Ergebnis beeinflussen können. Verschiedene Anwender sind z.B. Assistenten, die lediglich die Texte laden, die Parameter entsprechend eines Schemas einstellen und die Ergebnisse abspeichern; Extraktionsverantwortliche, die befugt sind Zwischenergebnisse zu überprüfen und zu modifizieren, oder auch Terminologen bzw. fachliche Experten, die für die Validierung zu Rate gezogen werden (vgl. Fleischmann 2006:65). Der gesamte Extraktionsprozess kann selbstverständlich auch von Einzelpersonen durchgeführt werden.

3.1.2.2 Repräsentation & Export der Ergebnisse

Direkt im Anschluss an die Identifikation werden die Ergebnisse angezeigt. Je nach TEP bieten sich nun verschiedene Möglichkeiten, die Ergebnisse direkt zu bearbeiten (um z.B. Tippfehler zu korrigieren, Wörter aus der Gemeinsprache auszusortieren), zu speichern bzw. zu exportieren (vgl. Zielinski & Ramírez-Safar 2005:4). Vor dem Export können die Ergebnisse teilweise auch gefiltert, geordnet, manuell sortiert und klassifiziert werden, z.B. durch den Extraktionsverantwortlichen, der etwa nach Themengebieten sortierte potentielle Termini direkt an die entsprechenden Terminologen bzw. Experten weiterleiten kann (vgl. Fleischmann 2006:54). Für verteilte Workflows ist es erforderlich, dass das entsprechende TEP eine Unterteilung in Teilaufgaben unterstützt. So ist z.B. eine Vorsortierung und Klassifizierung der Extraktionsergebnisse nicht möglich, wenn das gesamte Ergebnis direkt in eine Terminologiedatenbank exportiert wird. In diesem Fall beginnt sofort nach der Identifikation die Phase der Nachbearbeitung, die oft bereits nicht mehr im TEP ausgeführt wird (vgl. Fleischmann 2006:52).

Je nach unterstützten Datenformaten und verwendeter Software für die Extraktion und die Verwaltung der Terminologie kann der Export direkt oder über Zwischenformate realisiert werden (vgl. Zielinski & Ramírez-Safar 2005:5). Falls Zusatzinformationen angezeigt werden (z.B. grammatikalische Angaben, zugehöriges Dokument (Fundstelle), Kontexte), besteht teilweise auch die Möglichkeit, diese ebenfalls zu extrahieren (vgl. Saß 2004:34.).

3.1.3 Nachbearbeitungsphase

Im Rahmen der Nachbearbeitung müssen die Identifikationsergebnisse bearbeitet und validiert werden. Dies kann in zwei Stufen geschehen: zunächst erfolgt eine vorläufige Überprüfung der Ergebnisse durch einen Extraktionsverantwortlichen, der ein bereits bereinigtes Ergebnis an den zuständigen Terminologen bzw. fachlichen Experten zur Durchführung der

Validierung weiterleitet. Dieser entscheidet über die terminologische Relevanz des Terminus und somit über die Aufnahme in die Terminologiedatenbank (vgl. Fleischmann 2006:54). Es ist auch möglich die Validierung einem Terminologiezirkel mit mehreren Mitgliedern anzuvertrauen (vgl. König: 2009:60), oder eine Einzelperson sämtliche Aufgaben erledigen zu lassen.

Einige TEP ermöglichen nur eine eingeschränkte Weiterverarbeitung oder Verwendung, z.B. die Übersetzung der Termini und Verwendung der Terminologie als Grundlage für die nachfolgende Übersetzung. Eine dauerhafte Speicherung der Terminologie ist mit diesen Programmen nicht möglich (vgl. Zerfaß 2006:23). Durch die unterschiedlichen Möglichkeiten, die TEP bieten, gibt es Überschneidungen zwischen der Bearbeitung der Ergebnisse vor dem Export und der Weiterverarbeitung nach dem Export: einige Programme bieten zahlreiche Möglichkeiten zur Weiterverarbeitung bereits vor dem Export an (hier *interne Nachbearbeitung*), andere erlauben dies nicht, so dass die Weiterverarbeitung vollständig ausgelagert (z.B. in Tabellenverarbeitungsprogrammen oder in TVS) stattfindet (*externe Nachbearbeitung*) (vgl. Zielinski & Ramírez-Safar 2005:4).

3.1.3.1 Interne Nachbearbeitung

Falls das entsprechende TEP eine Bearbeitung erlaubt, können eventuell Zusatzinformationen (Beispielsätze, vermutete Beziehungen und Hierarchien zwischen einzelnen Termini, statistische Werte, etc.) gesucht (z.B. mittels KWIC-Suche), ausgewählt, editiert und gespeichert werden (vgl. Ahmad & Rogers 2001:749ff.). Diese sind sowohl für die Validierung als auch für die Terminologiedatenbank von Bedeutung. Sollte bereits validierte Terminologie vorhanden sein, bietet sich eventuell die Möglichkeit im Rahmen von Terminologieerkennung bestehende Terminologien abzugleichen bzw. über Negativlisten oder Ausschlussdatenbanken an dieser Stelle bereits zuvor abgelehnte potentielle Termini auszuschließen. Bei der bilingualen Terminologieextraktion können an dieser Stelle eventuell Terminologiepaare gesucht, editiert bzw. validiert werden (vgl. Saß 2004:54f.).

Für translatorische Zwecke kann im Falle von übersetzungsvorbereitenden Extraktionen die direkte Übersetzung der Identifikationsergebnisse angeboten werden, um im Laufe des anschließenden Übersetzungsprozesses bereits über ein zweisprachiges Glossar zu verfügen (vgl. Zerfaß 2006:23). Auch eine Auswahl von zusammengehörigen Terminologiepaaren und zugehörigen Kontextsätzen im Rahmen der übersetzungsnachbereitenden Extraktion für den anschließenden direkten Export in Terminologiedatenbanken kann an dieser Stelle erfolgen (vgl. Zerfaß 2006:22).

Im Anschluss an eine interne Nachbearbeitung erfolgt in vielen Fällen zusätzlich eine externe Nachbearbeitung, in deren Rahmen z.B. die Validierung der Termini und die Pflege der Terminologiedatenbank stattfinden.

3.1.3.2 Externe Nachbearbeitung

Wenn das entsprechende Programm keinerlei interne Nachbearbeitung ermöglicht, müssen erwünschte Zusatzinformationen nach dem Export hinzugefügt werden. Eventuell liefert das TEP automatisch ausgewählte Kontextinformationen (z.B. die Häufigkeit) mit, die andernfalls nur noch durch eine externe Durchsuchung des Korpus zu rekonstruieren sind (vgl. Fleischmann 2006:76).

Oft liegt die Hauptaufgabe der externen Nachbearbeitung in der Validierung der Termini, der Eingabe bzw. des Imports dieser in die Terminologiedatenbank sowie in der weiteren Pflege der Datenbank. Das Validieren ist eine nicht zu unterschätzende Teilaufgabe des Extraktionsprozesses: Die Identifikationsergebnisse sind lediglich potentielle Termini, die nicht ohne vorherige Überprüfung in Terminologiedatenbanken übernommen werden können (vgl. Ahmad & Rogers 2001:740). Bedeutend für die Validierung ist ein möglichst umfangreiches Informationsangebot, das eventuell auch in die Datenbank übernommen werden kann und somit auch die Pflege und Aktualisierung erleichtern kann (vgl. Ahmad & Rogers 2001: 740, Fleischmann 2006:54).

3.1.3.3 Regelmäßige Terminologieextraktion

Kompatibilität zwischen Datenbank und TEP können vor allem bei Anwendungen der regelmäßigen TE positive Auswirkungen auf die Qualität und Effizienz der Terminologieextraktion haben. Einige TEP bieten die Möglichkeit, die Terminologie nach der Ersterfassung mit jedem weiteren verfassten Text schrittweise zu erweitern, indem Varianten und Synonyme automatisch erkannt und zur Überprüfung vorgeschlagen werden. Die Terminologieextraktion selbst kann außerdem in weitere Prozesse integriert sein, welche im nächsten Kapitel betrachtet werden.

3.2 Weitere Anwendungsmöglichkeiten

Terminologieextraktionsprogramme können als Teil von Softwarepaketen in Prozesse zur Kontrolle der terminologischen Konsistenz bei der Verfassung von Ausgangstexten (vgl. Reuther & Rösener 2007:44) sowie zur Überprüfung der bilingualen Terminologie von Übersetzungen (vgl. Zerfaß 2008:12) integriert werden.

Die Konsistenzprüfung in der Ausgangssprache führt zu einheitlicher formulierten Texten mit einer einheitlichen Terminologie, die den Kosten- und Zeitaufwand einer Firma für

die Übersetzung ihrer firmeneigenen Dokumente deutlich reduzieren kann (vgl. Wohlfahrt 2008:5). Die Qualitätssicherung in der Übersetzung kann optional bzw. verpflichtend umgesetzt werden: In der optionalen Variante werden die jeweiligen Termini als Vorschlag in der Übersetzungsumgebung eingeblendet, in der verpflichtenden Variante wird eine Meldung für jede Übersetzung ausgegeben, die nicht entsprechend der in der Terminologiedatenbank gespeicherten Information formuliert wurde (vgl. Fleischmann 2006:62). Es kann auch im Anschluss an die Übersetzung automatisch ein detaillierter Report mit Verweisen auf Konsistenzmängel und Referenzmaterial erstellt werden, den der Übersetzer oder Lektor für die Korrektur der Übersetzung verwenden kann (vgl. Wohlfahrt 2008:7). Wiederum können in solchen Szenarien jeweils die bisher nicht bekannten Termini aus Originaltexten und Übersetzungen extrahiert und zur Erweiterung der Terminologiedatenbank verwendet werden (vgl. Wohlfahrt 2008:6f.). Falls die Texte vor der Konsitenzprüfung vollständig geparst werden, können auch Schreib- und Ableitungsvarianten, syntaktische Varianten sowie Synonymrelationen erkannt werden (vgl. Reuther & Rösener 2007:44).

Auch Programme anderer Art, z.B. wirtschaftlich ausgerichtete Programme, wie z.B. SAP, oder *Content Management Systeme* (CMS) können über Schnittstellen zu Terminologiedatenbanken verbunden werden. Eine derartige Verknüpfung kann die Terminologiekonsistenz erheblich fördern (vgl. Wohlfahrt 2008:6).

Die immer wieder erhoffte automatische Erstellung von Begriffsnetzen anhand der extrahierten potentiellen Termini bzw. die automatische Erfassung von Definitionen befinden sich laut Witschel noch in frühen Entwicklungsstadien, und liefern derzeit noch wenig Ergebnisse bzw. müssen durch aufwendige manuelle Tätigkeiten unterstützt werden (vgl. Witschel 2005:118 sowie Holz & Kranebitter 2010:20f.).

3.3 Probleme der computergestützten Terminologieextraktion

In diesem Kapitel sollen bei der Anwendung von TEP auftretende Probleme genauer betrachtet werden. In den meisten Publikationen wird die Qualität des Identifikationsergebnisses besonders eingehend thematisiert. Dementsprechend werden zunächst die Ursachen für eine geringe Qualität des Identifikationsergebnisses betrachtet. Im Anschluss wird auf die Probleme im Rahmen der Vorbereitung, Nachbearbeitung und im Zusammenhang mit der Terminologiearbeit sowie auf die mangelnde Vielseitigkeit der Programme eingegangen.

3.3.1 Qualität des Identifikationsergebnisses

Unterschiedliche Anwender haben unterschiedliche Vorstellungen davon, welche Wörter und Wortgruppen relevante Termini darstellen (vgl. Kapitel 1.1). Computer haben keinerlei Vorstellung davon, was Termini sind und welche relevant sein könnten. Zur Identifikation von

potentiellen Termini führen sie lediglich Maschinencode aus, in dem bestimmte Algorithmen implementiert wurden, die von Menschen anhand von Charakteristika für typische Termini entwickelt wurden (vgl. Kapitel 2.3) – sie können allerdings schnell und zuverlässig große Datenmengen verarbeiten (vgl. Ahmad & Rogers 2001:740).

Das Ergebnis hängt jedoch nicht nur von der Qualität des Programms (und der Definition der Charakteristika) ab, sondern auch vom Text bzw. Korpus und seinen Eigenschaften sowie vom Anwender und seinen Vorstellungen. Die Werte Silence und Noise bzw. Recall und Precision, die im Zusammenhang mit der Qualität der Terminologieidentifikation oft verwendet werden (siehe Kap. 2.4.1), sind dementsprechend ebenfalls abhängig von den Vorstellungen des Anwenders. Die Eignung eines Programms kann somit nicht pauschalisiert werden: je nach Situation, Texteigenschaften und Anwendervorstellungen können verschiedene Programme für unterschiedliche Situationen unterschiedlich gute Ergebnisse erzielen.

Auch nach einer erfolgreichen Anwendung muss weiterhin mit zahlreichen Fehlern und verzerrten Ergebnissen gerechnet werden, da zahlreiche Zusammenhänge, die ein Mensch beim manuellen Extrahieren bemerken würde, von den TEP nicht erkannt werden können (vgl. Ahmad & Rogers 2001:753). Die Überprüfung und Bereinigung der Ergebnisse im Rahmen der Nachbearbeitung ist daher in jedem Fall erforderlich (Ahmad & Rogers 2001:740, Zielinski & Ramírez-Safar 2006:28).

3.3.1.1 Vorverarbeitung

Im Bereich der Vorverarbeitung können bereits Fehler bei der *Wortsegmentierung* auftreten, indem Wörter bzw. Wortgruppen an den falschen Stellen geteilt bzw. nicht geteilt werden (vgl. Cabré *et al.* 2001:54). Vor allem die Identifikation von Mehrwortbenennungen kann durch die fehlerhafte Trennung von Token und Chunks stark beeinflusst werden (vgl. Ahmad & Rogers 2001:754f.). Außerdem können z.B. Regeln, die keine Zahlen oder Sonderzeichen innerhalb von Wörtern zulassen, die Identifikation einiger Termini von vornherein verhindern (vgl. Witschel 2005:94).

Cabré *et al.* weisen darauf hin, dass auch ein fehlerhaftes *POS-Tagging* für schlechte Ergebnisse verantwortlich sein kann (vgl. 2001:83). Durch der falschen Wortart zugeordnete Homographe kann nicht nur direkt das Identifikationsergebnis beeinflusst werden, indem die POS-Filter auf einer falschen Grundlage arbeiten, sondern auch eine fehlerhafte Grundformenbildung verursacht werden.

Eine fehlerhafte *Grundformenbildung* hat wiederum zur Folge, dass einige potentielle Termini nicht identifiziert werden können (vgl. Witschel 2005:94).

3.3.1.2 Identifikation

Entsprechend der verwendeten Identifikationsmethoden entstehen spezifische Probleme der einzelnen TEP. Zunächst sollen allgemeine Probleme bei der Identifikation potentieller Termini betrachtet werden, und in Folge soll auf die Probleme der linguistischen bzw. statistischen Methoden eingegangen werden.

Terminologieidentifikation beruht auf bestimmten Charakteristika wie der Häufigkeit im Text (siehe Kap. 2.4.3) oder der besonderen morphologischen bzw. morphosyntaktischen Struktur (siehe Kap. 2.4.4). TEP können nicht unterscheiden, ob ein potentieller Terminus zur Fachsprache oder zur Gemeinsprache gehört, und schon gar nicht, ob er terminologisch relevant ist (vgl. Cabré *et al.* 2001:54). Sie können lediglich Vorschläge präsentieren, die den Charakteristika entsprechen (vgl. Ahmad & Rogers 2001:740).

Die *Polysemie* stellt hier ein besonderes Problem dar, dass auch nicht durch Vergleichskorpora bzw. linguistische Vorverarbeitung gelöst werden kann: Ein TEP hat keine Möglichkeiten, zu erkennen, welchem Begriff polyseme Benennungen zuzuordnen sind (vgl. Ahmad & Rogers 2001:752), und kann sie dadurch nicht unterscheiden. Dies kann z.B. dazu führen, dass potentielle Termini aufgrund der Häufigkeit eines polysemen Begriffs in der Gemeinsprache nicht als potentielle Termini identifiziert werden, oder dass die Häufigkeiten unterschiedlicher Begriffe zu einer Häufigkeit (der Benennung) addiert werden und das Ergebnis verzerren.

Auch die *Variation* ist problematisch: TEP können nicht wissen, dass anderslautende oder geringfügig veränderte Benennungen sich auf den gleichen Begriff beziehen (vgl. Ahmad & Rogers 2001:753). Selbst Rechtschreibfehler können die erfolgreiche Identifikation von potentiellen Termini erschweren (vgl. Zerfaß 2006:23).

Die *Auslassung* von Wort- bzw. Wortgruppenbestandteilen, ein oft angewendetes Stilmittel zur Vermeidung von Wiederholungen, kann die Identifikation von Termini ebenfalls verhindern, da die Zusammengehörigkeit mit den ausgeschriebenen Benennungen für ein TEP nicht mehr erkennbar ist (vgl. Ahmad & Rogers 2001:753).

Statistische Verfahren

In Kapitel 2.4.3 wurde bereits darauf hingewiesen, dass für qualitativ hochwertige Ergebnisse umfangreiche Korpora notwendig sind. Die häufig verwendete Methode der Häufigkeit verursacht hohe Noise-Werte, die gegensteuernde Mindestfrequenz hingegen hohe Silence-Werte. Außerdem führt die in rein statistischen Verfahren fehlende Grundformenbildung ebenfalls zu hohen Noise-Werten (vgl. Kap. 2.4.3, Seite 23). Darüber hinaus werden in linguistisch nicht angereicherten Korpora Homographe mit verschiedenen Wortarten nicht ent-

sprechend ihrer Wortarten unterschieden und können ebenfalls das Ergebnis verzerren (vgl. Ahmad & Rogers 2001:753).

Ein generelles Problem der rein statistischen Ansätze sind potentielle Termini mit einer geringen Frequenz (*low-frequency-terms)*, die mit statistischen Methoden nicht identifiziert werden können und zu hohen Silence-Werten führen können (vgl. Cabré *et al*. 2001:55).

Linguistische Verfahren

Die starke Sprachgebundenheit und geringe Verfügbarkeit der TEP in anderen Sprachen als Englisch wurden bereits in Kapitel 2.4.4 erwähnt. Auch bei linguistischen Verfahren gibt es Ursachen für hohe Noise- bzw. Silence-Werte: Noise wird in linguistischen Verfahren vor allem durch zufällig nebeneinanderstehende Token verursacht, die den gesuchten POS-Mustern entsprechen ohne eine lexikalische Einheit zu bilden – Silence hingegen durch stark eingeschränkte Suchmuster, die es verhindern Mehrwortbenennungen zu identifizieren, die von diesen Mustern abweichen (vgl. Kap. 2.4.4, Seite 25).

Weitere Faktoren für erhöhte Silence-Werte bei der Identifikation von Mehrwortbenennungen sind die formale Variation (die Möglichkeit, Benennungen geringfügig umzuformulieren, ohne ihre Bedeutung zu verändern) und die Trennbarkeit (die Möglichkeit, Bestandteile von Mehrwortbenennungen an getrennten Positionen im Satz zu verwenden). Beide Phänomene verhindern eine Zuordnung des entsprechenden POS-Musters (vgl. Witschel 2005:26). Witschel kommt jedoch zu dem Schluss, dass diese beiden Phänomene eventuell vernachlässigbar sind (vgl. Witschel 2005:94f.).

Hybride Verfahren

Die Lösung der Probleme bei der Terminologieidentifikation scheint in hybriden Verfahren, in Kombinationen statistischer und linguistischer Methoden zu liegen und darüber hinaus in der Möglichkeit, semantische Informationen mit einzubeziehen (vgl. Cabré *et al.* 2001:84). Allerdings berichten auch jene Autoren, die hybride TEP betrachten von hohen Noise- bzw. Silence-Werten und einem dementsprechend hohen Nachbearbeitungsaufwand (vgl. Witschel 2005:96f.).

3.3.2 Vorbereitung

In der Vorbereitungsphase treten vor allem Kompatibilitätsprobleme auf. Diese können sowohl die Kompatibilität der Datenformate des Korpus als auch die Kompatibilität verschiedener Programme betreffen, die für den Extraktionsvorgang zusammenarbeiten sollten.

Zahlreiche TEP können lediglich Dateiformate verbreiteter Textverarbeitungsprogramme (z.B. Microsoft Word) einlesen. Problematisch gestalten sich Dateiformate, die zusätz-

liche Informationen in Form von Tags im Fließtext enthalten, z.B. multilinguale Dateiformate wie MLG- oder TMX-Dateien. TEP, die diese Informationen nicht als Tags erkennen, identifizieren sie unter Umständen als Termini (vgl. Gasser 2004:21-24). Falls die Tags über Stoppwortlisten herausgefiltert werden, und die Stoppwörter vom TEP zur Segmentierung verwendet werden, können sie Fehler bei der Segmentierung hervorrufen. Auch PDF-Dokumente müssen zunächst mithilfe von speziellen PDF-Konvertern bearbeitet werden (vgl. Fleischmann 2006:50f.).

Ein weiteres korpusbezogenes Problem, dass in der Vorbereitungsphase eventuell bedacht werden sollte, ist die unbekannte Leistungsfähigkeit der Programme. Der Anwender weiß in vielen Fällen nicht, ob das Programm dazu in der Lage ist, alle Dateien des Korpus in einem Zug zu verarbeiten und wie lange ein solcher Prozess dauern kann (vgl. Zerfaß 2006: 24, Saß 2004:69). Aus diesem Grund werden in der Regel Testdurchläufe vollzogen, in deren Rahmen auch bereits mögliche Einstellungen getestet werden können – dies sollte bei der Zeitplanung eines Extraktionsprojektes beachtet werden.

3.3.3 Nachbearbeitung

Die bei der Identifikation verwendeten Qualitätsmerkmale haben zweierlei Auswirkungen auf das Ergebnis: Noise führt zu einem umfangreichen Ergebnis, aus dem die nicht gewünschten potentiellen Termini entfernt werden müssen (vgl. Cabré *et al.* 2001:82). Dies geschieht in den meisten Fällen manuell, es sei denn, es liegen Non-Term-Listen vor, die eventuell zum Filtern verwendet werden könnten. Silence hingegen wird in praktischen Anwendungsszenarien nicht entdeckt: die potentiellen Termini bleiben unerkannt und werden nicht erfasst. Sie können nur durch eine vollständige manuelle Sichtung des gesamten Korpus gefunden werden, die sehr zeit- und kostenintensiv ausfallen kann (vgl. Thurmair 2003:175), und vermutlich nur in Evaluationssituationen zur Testung des Programmes erfolgt.

Einige Programme schränken die Nachbearbeitungsmöglichkeiten ein, indem sie lediglich den Export (vgl. Fleischmann 2006:52) oder die Übersetzung und unmittelbare Verwendung der potentiellen Termini ermöglichen (vgl. Zerfaß 2006:23).

Die fehlende Möglichkeit, Zusatzinformationen zusammen mit den potentiellen Termini zu exportieren, kann die Validierung erschweren, da Zusatzinformationen manuell gesucht werden müssten (vgl. Saß 2004:72, Ahmad & Rogers 749ff.).

3.3.4 Terminologiearbeit

Der Mangel an Zusatzinformationen kann neben der Validierung auch die Befüllung und vollständige Nutzung einer Terminologiedatenbank erschweren: Einer der bedeutendsten Vorteile von Terminologiedatenbanken im Vergleich zu reinen Wortlisten liegt in der Ver-

knüpfung der einzelnen Einträge untereinander. Synonyme, Wortfamilien, Themengruppen, Kontextbeispiele und weitere Informationen bereichern den Informationsgehalt von Terminologiedatenbanken (vgl. Fleischmann 2006:32ff.). Die Identifikationsergebnisse eines TEP werden in vielen Fällen nach der Häufigkeit geordnet oder alphabetisch sortiert und sind somit bei der Repräsentation nicht mehr in ihrem ursprünglichen Kontext enthalten. Falls die Programme keine zusätzlichen Informationen liefern bzw. sie nicht zum Export bereitstellen, müssen die Zusatzinformationen manuell erfasst und hinzugefügt werden (vgl. Saß 2004:72, Zerfaß 2006:22). Lieske stellt fest, dass TEP bisher wenig semantische bzw. konzeptionelle Informationen anbieten (vgl. Lieske 2002:130, vgl. auch Ahmad & Rogers 2001:749). Holz & Kranebitter bestätigen hingegen bereits brauchbare Extraktionen von Beziehungen zwischen verschiedenen potentiellen Termini (vgl. 2010:21).

Gasser versucht Informationen zu möglichen Synonymen über den Umweg der bilingualen Extraktion zu erhalten, indem sie all jene Termini als „synonymverdächtig“ kennzeichnet, die im bilingualen Ergebnis denselben fremdsprachigen potentiellen Termini zugeordnet wurden (vgl. Gasser 2004:34). Über die Effektivität dieser Methode ist nichts bekannt – begriffsorientierten Standards wird sie allerdings nicht gerecht, Fleischmann schlägt stattdessen die Aufgabenteilung im Rahmen eines verteilten Workflows vor (vgl. Fleischmann 2006:55f.).

Wie bereits in Kapitel 1.4.2 besprochen, ist die Begriffsorientierung ein wichtiges Charakteristikum der Terminologiearbeit. Dies bezieht sich nicht nur auf die systematische, sondern auch auf die punktuelle Terminologiearbeit, die vor allem bei der Erstellung von Terminologiedatenbanken für Firmen zur Anwendung kommt. Die Ordnung der Terminologie geht in der begriffsorientierten Terminologie vom Begriff aus, dies wird in der Eintragsstruktur von Terminologiedatenbanken oft über die Verwendung der Definition als Eintragsstichwort realisiert. Benennungen werden der Definition bzw. dem Begriff nur zugeordnet, sie sind nicht selbst die Grundlage des jeweiligen Eintrags (vgl. Fleischmann 2006:32ff.). Um Extraktionsergebnisse aus einem TEP in eine derartige Terminologiedatenbank zu extrahieren, muss die Eintragsstruktur der jeweiligen Programme kompatibel sein. Ist dies nicht der Fall müssen Übergangsschritte über Transferdatenbanken bzw. CSV-Dateien gewählt werden (vgl. Gasser 2004:7ff, Fleischmann 2006:83). Thurmair schlägt die Verwendung eines standardisierten Exportformats (OLIF) als Alternative zu CSV-Dateien vor (vgl. 2003:176), das allerdings von den TEP unterstützt werden müsste.

3.3.5 Vielseitigkeit der Programme

Zahlreiche Programme sind nur für eine beschränkte Anzahl an Sprachen (v.a. für Englisch) verfügbar (vgl. Cabré *et al.* 2001:82). Der Arbeitsaufwand für die Anpassung dieser Pro-

gramme für andere Sprachen erfordert bei linguistischen (und hybriden) Programmen in jedem Fall eine Veränderung der Programmes selbst, teilweise in Anbetracht der sprachlichen und grammatikalischen Unterschiede sogar eine vollständige Neukonzipierung[15] (vgl. Witschel 2005:99). Diese Anpassung kann keinesfalls von einem Anwender durchgeführt werden.

Statistische Programme können in zahlreichen Fällen für jede beliebige Sprache verwendet werden, jedoch muss auch bei diesen Verfahren Vorarbeit für die Verwendung einer neuen Sprache getätigt werden: z.B. die Erstellung oder Anpassung von Stoppwortlisten bzw. Trennzeichenlisten oder die Zusammenstellung von Vergleichskorpora (vgl. Cabré *et al.* 2001:79). Lieske und Wanhoff stellen fest, dass auch statistische Programme in verschiedenen Sprachen unterschiedlich gute Identifikationsergebnisse erzielen, dementsprechend auch nicht vollkommen sprachunabhängig sind (vgl. Lieske 2002:130, Wanhoff 2007:40). Darüber hinaus wurden einige TEP für spezielle Themenbereiche entwickelt und müssen für eine anderweitige Verwendung ebenfalls im Vorfeld angepasst werden (vgl. Cabré 2009:171f.).

3.4 Desiderata für die Weiterentwicklung

An Desiderata für die Entwicklung von TEP mangelt es in den verschiedensten Publikationen nicht. Ein Konsens besteht im Wunsch nach einer Verbesserung der Ergebnisqualität: möglichst wenig Noise bei möglichst geringen Silence-Werten (vgl. Heid 2003:189, Cabré *et al.* 2001:84), wobei ein gewisser Noise-Anteil notwendig zu sein scheint, um den unterschiedlichen Anforderungen der Anwender gerecht zu werden (vgl. Heid 2006:98). Darüber hinaus ist ein damit einhergehender verringerter Nachbearbeitungsaufwand erwünscht (vgl. Zielinski & Ramírez-Safar 2005:28). Den Weg dahin soll die Einbeziehung von semantischen Informationen bei der Entwicklung hybrider Verfahren weisen (vgl. Cabré *et al.* 2001:84).

Anwender wünschen außerdem mehr Zusatzinformationen (z.B. Kontexte, Definitionen, semantische Beziehungen) sowohl für die interne als auch für die externe Nachbearbeitung, und mehr Exportmöglichkeiten für diese Zusatzinformationen (vgl. Zielinski & Ramírez-Safar 2005:28) sowie die Verwendung von offenen, standardisierten Formaten für den Export (vgl. Thurmair 2003:175f.). Cabré *et al.* fordern darüber hinaus die Einbeziehung von Verben und Adjektiven (vgl. 2001:84). Dies entspricht auch den Arbeitsbedingungen zahlreicher Anwender, die in einer Onlineumfrage zu Terminologieextraktionsprogrammen zu 86% angaben, auch Verben in ihre Terminologie aufzunehmen (vgl. Zielinski & Ramírez-Safar 2006:26).

[15] Die Anpassung eines komplexen hybriden Programms von der englischen auf die deutsche Sprache beschreibt Witschel (vgl. 2005:99-109).

Ein weiterer Wunsch besteht in der Verfügbarkeit von linguistischen bzw. hybriden Programmen für weitere Sprachen (vgl. Cabré *et al.* 2001:84), zahlreiche Anwender arbeiten mit drei bis vier verschiedenen Sprachen und können die computergestützte Extraktion nur mithilfe von statistischen Programmen in allen Sprachen anwenden (vgl. Zielinski & Ramírez-Safar 2005:27).

3.5 Zusammenfassung

Im Verlauf dieses Kapitels wurden verschiedene Aspekte der computergestützten Terminologieextraktion in der Praxis besprochen. Zunächst wurden die Phasen der Anwendung eines TEP beschrieben. Die einzelnen Aufgaben und Arbeitsschritte der beteiligten Akteure sowie die verschiedenen Möglichkeiten, die die Programme zur Unterstützung des Prozesses bieten, wurden vorgestellt. Anschließend wurden weitere Anwendungsmöglichkeiten, vor allem im Sinne der Konsistenzprüfung, präsentiert. Ein Schwerpunkt des Kapitels beschäftigt sich mit den Problemen der computergestützten TE und betrachtet sowohl ergebnisbezogene als auch prozessbezogene Faktoren. Den Abschluss bildet eine Zusammenfassung der Wünsche für weitere Entwicklungen von Terminologieextraktionsprogrammen.

Das nächste Kapitel erarbeitet die Grundlagen der Evaluation von sprachverarbeitenden Programmen und erläutert im Detail die Entwicklung einer Evaluationsmethodologie anhand des EAGLES-Evaluationsrahmens.

4 Evaluation von sprachverarbeitenden Programmen

Dieses Kapitel beschäftigt sich mit den methodologischen Grundlagen der Softwareevaluation. Zunächst wird ein allgemeiner Einblick in die Grundlagen der Softwareevaluation und Softwarequalität gegeben (Kap. 4.1). Anschließend wird die Struktur des EAGLES-Evaluationsrahmen (Kap. 4.2) und die Entwicklung von Anforderungen, Merkmalen und Maßen (Kap. 4.3) skizziert. Das Design einer Evaluationsmethodologie anhand des EAGLES-Evaluationsrahmen wird in Kapitel 4.4 betrachtet und in Kapitel 4.5 werden die einzelnen Schritte vom Beginn der Planung bis zur Durchführung einer Evaluation in einem Ablauf zusammengefasst.

4.1 Konzepte, Grundlagen & Begriffe

Ziel dieser Arbeit ist es nach einer detaillierten Betrachtung der computergestützten Terminologieextraktion, bisherige Evaluationen von TEP zu betrachten und Ideen für erweiterte methodologische Grundlagen für die Evaluation von Terminologieextraktionsprogrammen zu entwickeln. Terminologieextraktionsprogramme gehören neben zahlreichen weiteren Programmtypen zu den sprachverarbeitenden Programmen (*NLP*[16] *systems*). Die Literatur dieses übergeordneten Bereichs soll in diesem Kapitel eingehend behandelt werden und die Grundlage für die anschließende Betrachtung der Evaluationen von TEP (siehe Kap. 5) bilden. Der Bezug zu TEP wird auch in diesem Kapitel beibehalten, Beispiele werden soweit möglich an die Terminologieextraktion angepasst.

In einigen Arbeiten im Bereich der Terminologieextraktion wurden bereits Evaluierungen durchgeführt bzw. Evaluationsmethodologien vorgestellt. Diese konzentrieren sich jedoch oft auf einen einzelnen Aspekt: die quantitative Betrachtung der Identifikationsergebnisse (siehe Kap. 5.2). Einige Arbeiten bemühen sich, ganzheitlichere Evaluationen zu planen und durchzuführen, diese werden in Kapitel 5.3 vorgestellt. Als Grundlage für umfassendere Evaluationen dient die Arbeit der EAGLES-Gruppe zur Schaffung eines gemeinsamen Rahmens für die Evaluation von sprachverarbeitenden Programmen. Auch die Ideen zur Evaluationsmethodologie dieser Arbeit sollen auf den Grundlagen des EAGLES-Rahmens aufbauen.

In früheren Arbeiten wurde der Bezug zu EAGLES zwar festgestellt, die EAGLES-Grundlagen wurden aber im Rahmen der Arbeiten kaum explizit thematisiert. Die Vorgangsweise zur Erarbeitung der Methodologie wurde dementsprechend in manchen Fällen nicht wirklich deutlich (vgl. z.B. Saß 2004). Da eine derartige Entwicklung in dieser Arbeit verhindert werden soll, widmet sich dieses Kapitel zunächst ausführlich den Grundlagen der Softwareevaluation (siehe Kap. 4.1.1 und 4.1.2) mit einem besonderen Schwerpunkt auf dem

[16] NLP steht hier für *natural language processing* (vgl. Sparck-Jones & Galliers 1995:1)

EAGLES-Evaluationsrahmen (4.2 - 4.5) und auf ergänzender Literatur zur Evaluation von Programmen, insbesondere sprachverarbeitenden Programmen.

Deutsche Terminologie

Bevor in die Evaluationsthematik eingestiegen werden kann, soll hier noch kurz ein Hinweis zur verwendeten deutschen Terminologie erfolgen. Ein Großteil der Publikationen zur Softwareevaluation wird auf Englisch verfasst. In deutschen Publikationen werden oft die englischen Bezeichnungen innerhalb des deutschen Textes verwendet, oder je nach Gebrauch eigenhändig übersetzt. Es soll nicht Ziel dieser Arbeit sein, die deutsche Terminologie der Softwareevaluation zu beschreiben bzw. vereinheitlichen. Dementsprechend wird es hierzu keine derart ausführliche Diskussion der Begriffe und Bezeichnungen wie im Fall der Terminologie zur Terminologieextraktion geben (vgl. Kapitel 1.3). Für Konzepte, die aus englischsprachigen Quellen zitiert werden, wird jeweils der Verweis auf ein deutschsprachiges Dokument angegeben, in dem dasselbe Konzept oder Teile dieses Konzepts in deutscher Sprache behandelt wurden. Die Begriffe werden aus diesen deutschsprachigen Dokumenten übernommen, und im weiteren Verlauf der Arbeit wird die erste eingeführte Benennung beibehalten. Es kann nicht ausgeschlossen werden, dass in anderen Publikationen andere Benennungen für dieselben Begriffe verwendet werden.

In einigen Fällen werden die Benennungen direkt aus dem Englischen übersetzt – in diesen Fällen wird hinter der ersten Nennung der jeweiligen Benennung die englische Bezeichnung angeführt. Für jene englischen Bezeichnungen, die auch in deutschen Publikationen konsequent in der englischen Version verwendet werden, wurde keine neue deutsche Bezeichnung geschaffen.

Das EAGLES-Projekt

Das EAGLES-Projekt (*Expert Advisory Group on Language Engineering Standards*) wurde von der Europäischen Kommission initiiert und gefördert und von Mitarbeitern und Forschern aus verschiedenen europäischen Universitäten und Forschungsinstitutionen durchgeführt (vgl. EAGLES 1995:1-4). Der Durchführungszeitraum des EAGLES-Projekts liegt bereits einige Jahre zurück (1993-1999). Nichtsdestotrotz werden die im Rahmen des Projektes entstandenen Dokumente aufgrund ihres grundlegenden Charakters auch in aktuellen Arbeiten zitiert (vgl. Jekat & Schultz 2004:574) und zur Entwicklung von Evaluationsmethodologien verwendet (vgl. Estrella *et al.* 2009).

Ziel der im Projekt durchgeführten Aktivitäten war vor allem ein größeres Maß an Standardisierung bei der Weiterentwicklung von sprachverarbeitenden Programmen. Einer der Projektzweige beschäftigte sich mit der Weiterentwicklung und Standardisierung der

Evaluation von sprachverarbeitenden Programmen. Dieses Kapitel beruht zu großen Teilen auf dem Abschlussdokument dieses Projektzweigs: *Evaluation of Natural Language Processing Systems* (EAGLES 1995[17]). In diesem Dokument wird ein Evaluationsrahmen skizziert, der für die Evaluation verschiedenster sprachverarbeitender Programme verwendet werden kann (vgl. EAGLES 1995:9f.). Ziel des Projektzweigs war es, Evaluationskriterien und zugehörige Testtechniken sowie generelle Richtlinien zu entwickeln, die es zukünftigen Testern ermöglichen, aus diesen Bausteinen innerhalb des Evaluationsrahmens ihr eigenes Design zusammen zu stellen (vgl. EAGLES 1995:9). Die Entwicklung dieses Evaluationsrahmens wird nicht als abgeschlossenes Projekt, sondern als Prozess angesehen, der auch nach Beendigung des Projektes fortgesetzt wird. Die Grundlagen des Evaluationsrahmens wurden anhand einzelner Programmtypen für einige Bereiche (Rechtschreib- und Grammatikprüfung sowie CAT-Programme) beispielhaft weiterentwickelt.

Das EAGLES-Projekt widmete sich einer spezifischen Art der Evaluation – der anwenderorientierten Adäquatheitsevaluation[18] im Sinne des *Consumer Report Paradigm*: eine möglichst neutrale Evaluation, die einem potentiellen Käufer/Anwender eines sprachverarbeitenden Programmes Entscheidungshilfen bietet – und verwendete hierfür als Ausgangspunkt die ISO-IEC Norm 9126 (vgl. EAGLES 1995:13, 40f.).

4.1.1 Grundlagen der Softwareevaluation

Bevor näher auf den EAGLES-Evaluationsrahmen eingegangen wird, soll zunächst das Konzept Evaluation genauer betrachtet werden. Durch eine Evaluation wird festgestellt, welchen Wert ein bestimmtes Objekt für eine bestimmte Person bzw. Personengruppe hat (vgl. EAGLES 1995:15). Um Objekte zu evaluieren, müssen die Anforderungen der Person (gewünschte Eigenschaften) und die Charakteristika des Objekts sowie weitere relevante Faktoren bekannt sein bzw. in Erfahrung gebracht werden.

Die evaluierten Objekte im Bereich der Softwareevaluation sind Programme bzw. Komponenten von Programmen. Die Personen, für die die Evaluation durchgeführt wird, können Softwareentwickler, potentielle Käufer bzw. potentielle Anwender sein. Diese beiden letzten Personengruppen sind nicht unbedingt deckungsgleich, da der Käufer eines Programms dieses nicht in jedem Fall benutzt: es kann z.B. auch für seinen Mitarbeiter gedacht sein und dieser kann durchaus andere Anforderungen an das Programm haben als der tatsächliche Käufer (vgl. EAGLES 1995:150). Darüber hinaus ist zu beachten, dass jene Personen,

[17] Eine aktuellere Version des Reports ist leider nicht erhältlich, lediglich eine sehr knapp gehaltene Sieben-Schritt-Anleitung aus dem Jahre 1999 (vgl. EAGLES 1999).

[18] Zahlreiche deutsche Bezeichnungen der EAGLES-Terminologie wurden dem Überblicksartikel *Evaluation sprachverarbeitender Systeme* (Jekat & Schultz 2004) entnommen.

für die eine Evaluation durchgeführt wird – die Auftraggeber –, andere sein können als jene, die an der Evaluation beteiligt sind: professionelle Evaluatoren können Evaluationen für potentielle Käufer durchführen und an Evaluationen für Softwareentwickler können potentielle Anwender beteiligt sein (vgl. EAGLES 1995:33 & 71).

4.1.1.1 Evaluationstypen

Im Softwarebereich können drei Typen der Evaluation unterschieden werden: Diagnoseevaluation (*diagnostic evaluation*), Fortschrittsevaluation (*progress evaluation*) und Adäquatheitsevaluation (*adequacy evaluation*). Die ersten beiden Formen sind v.a. für Softwareentwickler von Bedeutung, sie dienen der Systemverbesserung und werden angewendet, um ein bestimmtes Programm bzw. eine Komponente zu überprüfen sowie Fehler und Möglichkeiten zur Verbesserung zu entdecken (vgl. Sparck-Jones & Galliers 1995:157). Entwicklerorientierte Evaluationen werden weder im EAGLES-Bericht noch in dieser Arbeit näher behandelt. Adäquatheitsevaluationen werden angewendet, um verschiedene Programme zu vergleichen, und anhand der Ergebnisse das wertvollere, besser geeignete auszuwählen (vgl. EAGLES 1995:24). Da sie vor allem für potentielle Anwender interessant sind und sich nach ihren Bedürfnissen richten, werden sie auch als anwenderorientierte Evaluationen bezeichnet (vgl. Jekat & Schultz 2004:573f.) und werden den zentralen Evaluationstyp dieser Arbeit darstellen.

Darüber hinaus werden *Glass-Box* von *Black-Box*-Evaluationen unterschieden. Glass-Box-Evaluationen setzen Kenntnisse und Einblicke in die einzelnen Komponenten der Programme voraus, sind dementsprechend auch v.a. Entwicklern vorbehalten (vgl. EAGLES 1995:68). Da für potentielle Anwender von TEP nur im Fall einer eigens für den Anwender angefertigten Software die Möglichkeit besteht, Glass-Box-Evaluationen durchzuführen, soll diese Art der Evaluation in dieser Arbeit ebenfalls nicht behandelt werden. Stattdessen wird sich diese Arbeit auf Black-Box-Evaluationen konzentrieren.

King identifizierte in den 90er Jahren zwei unterschiedliche Ansätze der Softwareevaluation, die sehr gegensätzliche Positionen vertraten: mithilfe des bereits in den 50er Jahren entstandenen Ansatzes wurden Evaluationen auf Auftrag eines bestimmten Kunden für ein bestimmtes Produkt durchgeführt, das unter spezifischen Bedingungen angewendet werden sollte – die Überprüfung der Adäquatheit eines bestimmten Programms mit besonderer Berücksichtigung des potentiellen Kontextes stand hier im Vordergrund (vgl. King 2005:45f.).

Der zweite Ansatz wollte die Entwicklung von standardisierten Evaluationsmethodologien vorantreiben und sich zu diesem Zweck vollständig auf die Funktionalität der Programme konzentrieren – der Kontext und die internen Prozesse wurde im Rahmen derartiger Evaluationen nicht beachtet, lediglich die Ergebnisse der Programme wurden evaluiert (vgl.

King 2005:46ff.). Durch diese Fokussierung wurde den verwendeten Maßen mehr Aufmerksamkeit gewidmet: es mussten Maße entwickelt werden, die es ermöglichten unterschiedlich strukturierte Programme in Black-Box-Evaluationen zu vergleichen (vgl. King 2005:48).

Das EAGLES-Projekt hatte sich zum Ziel gesetzt, beide Ansätze zu vereinen (vgl. King 2005:49).

4.1.1.2 Evaluationsprozess

Eine Evaluation besteht aus mehreren Phasen, in denen die Vorbereitung, Durchführung und Auswertung der Evaluation erfolgt. Zum Zwecke der Standardisierung ist ein Evaluationsprozess in der ISO-IEC Norm 14598[19] beschrieben:

- Definition der Evaluationsanforderungen (Zweck der Evaluation, Produkttypen und Qualitätsmodell festlegen)
- Spezifizierung der Evaluation (Metriken auswählen, Grenzwerte einer Bewertungsskala für Metriken sowie Gewichtung der einzelnen Metriken festlegen)
- Design der Evaluation (Evaluationsplan erstellen)
- Durchführen der Evaluation (Messen, Ergebnisse mit Kriterien vergleichen und entsprechend der Gewichtung zusammenfassen)

(vgl. Gruber 2007:22, EAGLES 1995:65).

An diesem Modell wird deutlich, dass die Planung bei der Evaluation von besonderer Bedeutung ist: 75% des Prozesses werden in der Übersicht der Vorbereitung und Planung gewidmet. Darüber hinaus sind die Definition der Ziele sowie die Beschränkung der Evaluationsobjekte und die Festlegung der Evaluationskriterien bereits in der ersten Phase vorgesehen. Im Anschluss werden genaue Metriken und der Umgang mit diesen Metriken definiert, und schließlich eine detaillierte Planung der einzelnen Schritte entworfen. Erst im Anschluss kann die Evaluation durchgeführt werden, bei der vorher festgelegte Grenzwerte und Gewichtungen eingehalten werden müssen.

Jedoch muss beachtet werden, dass es keine derart einheitliche Herangehensweise an eine Evaluation gibt, wie diese Norm glauben lässt, sondern lediglich Empfehlungen bzw. Richtlinien oder bewährte Praktiken (vgl. Sparck-Jones & Galliers 1995:124). Jede Evaluation unterscheidet sich von anderen Evaluationen, da jede Evaluation an die Evaluationsziele, die spezifischen Anforderungen und das Evaluationsobjekt angepasst wird und darüber hinaus von Faktoren wie vorhandenen finanziellen Mitteln, verfügbarer Zeit und verfügbarem Personal beeinflusst wird (vgl. EAGLES 1995:73). Es können zwar allgemeine Phasen im Evaluationsprozess definiert werden, die in den meisten Evaluationen durchlaufen werden,

[19] Die ISO-IEC Norm 14598 wird derzeit überarbeitet. Die neuen Vorgaben der ISO-IC 25000 Gruppe konnten für diese Arbeit noch nicht berücksichtigt werden.

doch werden auch diese in einzelnen Publikationen von der Norm abweichend eingeteilt (vgl. Jekat & Schultz 2004:574). Im EAGLES-Modell, das sich in den Grundzügen an der Norm orientiert, werden zunächst die Anwendergruppen analysiert und ihre Anforderungen definiert, um dann anhand der Anforderungen Merkmale[20] zu entwickeln und mit Maßen zu versehen, denen schließlich Methoden zugeordnet werden (vgl. EAGLES 1995:20f.). Dieser Prozess wird in den Kapiteln 4.3 und 4.4 im Detail beschrieben.

4.1.1.3 Evaluationskriterien

Die wichtigste Grundlage jeder Evaluation sind die Evaluationskriterien. Diese sollten bei jeder Evaluation anhand der Charakteristika der jeweiligen Evaluationsobjekte und der Anforderungen der Auftraggeber der Evaluation (in diesem Fall Anwender bzw. Käufer) neu entwickelt bzw. angepasst werden (vgl. Jekat & Schultz 2004:576). Eine Evaluation ist immer in irgendeiner Form mit Qualität verbunden und im Sinne der Standardisierung gibt es auch für Qualitätskriterien Vorgaben in Form von Qualitätsmodellen in ISO-IEC-Normen (vgl. Kapitel 4.1.2). Diese dienen jedoch lediglich als Ausgangspunkt, da im Rahmen der Vorbereitung viel detailliertere Kriterien entwickelt werden müssen, als in den Normen vorgegeben werden können (vgl. Plösch 2009:6 & 11). Gemäß der EAGLES-Vorgaben sollten die Werte von Merkmalen beobachtbar, messbar oder direkt von den Werten anderer Merkmale ableitbar sein, adäquat alle expliziten und impliziten Anforderungen des Anwenders widerspiegeln und allgemein genug sein, um bei der Evaluation unterschiedlicher Programme bzw. für unterschiedliche Anwender(-gruppen) angewendet werden zu können (vgl. EAGLES 1995:20f.). Mit dem Prozess der Entwicklung dieser Merkmale beschäftigt sich das Kapitel 4.3.3.

4.1.1.4 Standardisierung von Evaluationen

Nicht nur auf dem Gebiet der Softwareevaluation wird eine immer stärkere Standardisierung angestrebt. Die Gründe für ein solches Ansinnen liegen auf der Hand: nur durch eine stärkere Standardisierung kann eine größere Vereinheitlichung und somit leichtere Vergleichbarkeit von Evaluationsergebnissen erreicht werden: „Qualität [soll] transparent und messbar [gemacht werden]“ (vgl. Plösch 2009:5). Auch der Austausch und das gemeinsame Verwenden

[20] Die Bezeichnung *Merkmal* wird in dieser Arbeit für das im EAGLES-Report verwendete *attribute* benutzt, und bezieht sich somit auf den EAGLES-Begriff. Mit der Bezeichnung *Kriterium* wird die englische Bezeichnung *feature* wiedergegeben (vgl. Kapitel 4.2). Darüber hinaus wird die Bezeichnung Evaluationskriterium für eine allgemeine Erwähnung des Konzepts Kriterium verwendet, die keiner spezifischen Terminologie zugeordnet werden kann.

von Methodologien, einzelnen Methoden und Testmaterialien würden durch eine stärkere Vereinheitlichung erleichtert (vgl. EAGLES 1995:38).

Jedoch müssen auch die Grenzen der Standardisierung beachtet werden: jede Evaluation ist ein eigenständiger Prozess mit spezifischen Prioritäten, Bedingungen und Einschränkungen – eine zu starke Standardisierung würde keinen Freiraum mehr für angepasste Evaluationskriterien und -methoden lassen (vgl. EAGLES 1995:18, 56 & 97). Teilweise erscheint eine Standardisierung bei der Kategorisierung und Entwicklung von Testtypen, sowie bei der Erstellung von Instrumenten als möglich (vgl. EAGLES 1995:98). Darüber hinaus bietet es sich an, nach einer genauen Analyse der spezifischen Evaluationssituation, einzelne Merkmale und Methoden aus ähnlichen Evaluationssituationen zu übertragen (vgl. Sparck-Jones & Galliers 1995:138).

4.1.2 Grundlagen der Softwarequalität

In diesem Kapitel sollen die für diese Arbeit relevanten Grundlagen der Softwarequalität betrachtet werden – eine umfassende Vorstellung des Bereichs Softwarequalität wird nicht angestrebt, lediglich ausgewählte Themen werden besprochen.

Im Rahmen einer Evaluation wird ein bestimmtes Programm für einen bestimmten Anwender bzw. Käufer bewertet. Der Wert des Programms hängt stark von der Qualität des Programms ab. Für eine genauere Betrachtung des Konzeptes Qualität soll hier das von der ISO vorgeschlagene Qualitätsmodell für Software vorgestellt werden. Der Oberbegriff Qualität wird in der ISO-IEC Norm 9126[21] in unterschiedliche Qualitätstypen unterteilt (interne Qualität, externe Qualität und Gebrauchsqualität). Die interne Qualität befasst sich mit vollständig programmabhängigen Faktoren (z.B. Programmiercharakteristika), die externe Qualität mit Faktoren der jeweils gegebenen programmunabhängigen Bedingungen (z.B. Hardware bzw. Testdaten). Mit der Gebrauchsqualität wird vor allem die Produktivität der Anwender gemessen (vgl. Bøegh 2008:59), sie kann nur in einer realen Anwenderumgebung überprüft werden (vgl. King 2005:50).

Die interne und externe Qualität werden weiter differenziert in

- Funktionalität[22] (Angemessenheit, Richtigkeit, Interoperabilität, Sicherheit und Ordnungsmäßigkeit/Konformität)
- Zuverlässigkeit (Reife, Fehlertoleranz, Wiederherstellbarkeit, Ordnungsmäßigkeit/Konformität)
- Benutzbarkeit (Verständlichkeit, Erlernbarkeit, Bedienbarkeit, Attraktivität, Ordnungsmäßigkeit/Konformität)
- Effizienz (Zeitverhalten, Verbrauchsverhalten, Ordnungsmäßigkeit/Konformität)

[21] Die ISO-IEC Norm 9126 bzw. ihr Inhalt werden voraussichtlich auch in der neuen Normgruppe ISO-IEC 25000 erhalten (vgl. Bøegh 2008:58), so dass das Qualitätsmodell auch weiterhin gelten wird.

[22] Die deutsche Terminologie wurde Neukirchen 2007:2ff. entnommen.

- Änderbarkeit (Modifizierbarkeit, Stabilität, Prüfbarkeit, Ordnungsmäßigkeit/Konformität)
- Übertragbarkeit (Anpassbarkeit, Installierbarkeit, Koexistenz, Austauschbarkeit, Ordnungsmäßigkeit/Konformität).

Die Gebrauchsqualität wird unterteilt in

- Effizienz
- Produktivität
- Sicherheit
- Zufriedenstellung

(vgl. Bøegh 2008:58).

Die Kriterien werden in der Norm jeweils kurz erklärt bzw. definiert (vgl. EAGLES 1995:59-62). Wie bereits in Kapitel 4.1.1.3 festgestellt sind jedoch auch diese Definitionen sehr weit gefasst und müssen für jeden Anwendungsfall ausgewählt und spezifiziert werden. Diese Spezifizierung kann sehr unterschiedlich ausfallen – es gilt dementsprechend auch in diesem Fall, dass trotz der angestrebten Standardisierung, ein Maßstab für Softwarequalität nicht allgemein festgelegt werden kann: Qualität als Konzept ist abhängig von den Bedürfnissen und Wünschen des Anwenders bzw. Käufers (vgl. EAGLES 1995:94). Es ist die Aufgabe des Evaluators, die Kriterien auszuwählen und je nach Bedürfnissen des Anwenders/Käufers zu gewichten (vgl. King 2005:50).

Aus diesem Grund wird bei der Entwicklung evaluationsspezifischer Kriterien im Rahmen von EAGLES ein wichtiges Element der Qualitätsmanagementsnorm ISO 8402, das auch in der DIN 8402 enthalten war, in den Vordergrund gestellt: „Qualität[23] ist die Gesamtheit von Merkmalen einer Einheit bezüglich ihrer Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen“ (vgl. EAGLES 1995:63). Die Qualitätsnorm wurde inzwischen durch die DIN EN ISO 9001:2000 abgelöst, in ihr wird Qualität definiert als „Grad, in dem ein Satz inhärenter Merkmale Anforderungen erfüllt“ (vgl. Pfitzinger 2002:5). Pfitzinger weist jedoch darauf hin, dass sich die zugehörige Erklärung kaum von der vorher gültigen Definition unterscheidet (vgl. Pfitzinger 2002:5). Der Schwerpunkt des Konzeptes Qualität liegt demnach nach wie vor in der Erfüllung von im Vorfeld festgelegten, d.h. expliziten, und vorausgesetzten, d.h. impliziten, Anforderungen bzw. Erfordernissen. Diese Anforderungen müssen jedoch zunächst definiert werden, auch für den Fall, dass sie als vorausgesetzt angesehen werden. Für spezielle Programme sind potentielle Anwender bzw. Käufer oft nicht in der Lage, ihre Anforderungen selbstständig zu spezifizieren (vgl. Bøegh 2008:59). Dies muss bei der Erfassung der Anforderungen durch den Evaluator oder auch Anwender beachtet werden (s.a. Kapitel 4.3.2).

[23] Die deutsche Formulierung wurde Hoffmann 2008:6 entnommen.

Neben der Unterscheidung von Gebrauchsqualität und interner sowie externer Qualität können auch noch die Konzepte technische Qualität und ökonomische Qualität unterschieden werden. Für die anwenderorientierte Adäquatheitsevaluation sind grundsätzlich alle Bereiche der Qualität interessant, die technische jedoch nur, insofern sie direkte Auswirkungen auf die Verwendung bzw. Änderbarkeit oder Übertragbarkeit des Programmes hat. Die ökonomische Qualität ist auch als Preisleistungsverhältnis bekannt, und in Kombination mit der Gebrauchsqualität, sowie deren Teilkriterium Produktivität, besonders für potentielle Käufer interessant (vgl. EAGLES 1995:40).

4.2 Der EAGLES-Evaluationsrahmen

Bevor der EAGLES-Evaluationsrahmen vorgestellt werden kann, müssen zunächst einige grundlegende Begriffe und Zusammenhänge aus dem Bereich der (sprachverarbeitenden) Software eingeführt werden.

Grundlagen

Ein *Programm* (*system*) besteht aus mehreren *Komponenten (components)*. Sobald ein Programm auf einem (oder mehreren) Computer(n) installiert ist, und somit in eine *Umgebung (environment)* integriert ist, kann von einem *Setup (setup)* gesprochen werden. Dieses Setup umfasst das installierte Programm mitsamt der Umgebung und den *Umgebungsfaktoren (environment factors)*. Zu den Umgebungsfaktoren gehören z.B. der *Datenfluss* zum/vom Programm und der bzw. die *Anwender (user)* des Programms. Ein Programm kann verschiedene *Parameter (parameter)* haben, die eingestellt werden können. Ein Anwender nimmt bei der Verwendung von Programmen bestimmte *Rollen (role)* an und führt *Tätigkeiten (activity)* durch (vgl. Sparck-Jones & Galliers 1995:8,11-14 & 23)[24].

In Folge werden nun die Grundlagen des EAGLES-Evaluationsrahmens zusammengefasst.

Struktur des EAGLES-Evaluationsrahmen

Programme verfügen über *Eigenschaften*[25] (*machine attributes*) und werden nach Programmtypen klassifiziert, *Anwender* (*user*) verfügen über *Interessen (interests)* und werden in *Anwendergruppen (user groups)* mit ähnlichen Interessen zusammengefasst. Anhand dieser Anwendergruppen können *Anwenderprofile* mit den *Anforderungen (requirements)* der Anwender erstellt werden *(user profiling)*. Für eine Evaluation werden zunächst ebendiese

[24] Bei den deutschen Begriffen handelt es sich in diesem Fall um direkte Übersetzungen. Für eine Übersicht dieser und weiterer Begriffe vgl. Sparck-Jones & Galliers 1995:16.

[25] Die Terminologie dieses Unterkapitels ist zu großen Teilen Lieske 2002 entnommen.

Anforderungen analysiert. Die Anforderungen können von den potentiellen *Anwendern (user)* oder *Käufern (customer)* explizit gestellt werden, bzw. vom *Evaluator (evaluator)* als implizit vorhanden identifiziert werden (vgl. EAGLES 1995:21-29).

Anhand der Anforderungen von Anwendern und Käufern und der Eigenschaften von Programmen werden *Merkmale (attributes)* entwickelt, die mit einer *Metrik (metric)* bzw. einem *Maß (measure)* versehen werden. Um ein Maß bei einer Evaluation zu verwenden, muss eine *Methode (method)* für das jeweilige Merkmal mit dem entsprechenden Maß entwickelt werden. Mithilfe der Methode werden dem Maß im Verlauf der Evaluation, z.B. durch *Messungen (measurement) Werte (values)* zugeordnet. Unterschiedliche Merkmale können unterschiedliche Wertetypen erfordern und werden deshalb nach Merkmaltypen klassifiziert (vgl. EAGLES 1995:22-31). Ein Merkmal mit seinem zugehörigen Wert wird auch als *Kriterium (feature)* bezeichnet (vgl. EAGLES 1995:14).

Anforderungen, Kriterien und Methoden können selbst definiert bzw. aus vorherigen Evaluationen übernommen oder aus einer *Bibliothek (library)* an Merkmalen mit zugeordneten Metriken und Methoden ausgesucht werden. Wichtig für den Evaluationsprozess ist es, dass sowohl Anforderungen als auch Merkmale, Kriterien und Methoden *gültig* (*valid*) und *zuverlässig* (*reliable*) sind (vgl. EAGLES 1995:29ff.).

Für eine Evaluation können beliebige Methoden miteinander kombiniert werden. Das Evaluationsdesign ist von zahlreichen Faktoren abhängig und beinhaltet die Auswahl der oder des *Testtyps (test type*), der verwendeten *Test- und Berichtsinstrumente* (*testing und reporting instruments*) und der verwendeten *Testmaterialien (test materials)* (vgl. EAGLES 1995:32-38).

Das EAGLES-Projekt strebt darüber hinaus die Schaffung einer *parametrisierbaren Testumgebung (parameterisable test bed)* für die Evaluation sprachverarbeitender Programme sowie die Entwicklung umfassender Bibliotheken zum Zweck einer besser unterstützten Weiterentwicklung sowie Standardisierung der Testmethoden an (vgl. EAGLES 1995:18ff.). King weist auf die Grenzen hin, die solchen Bibliotheken zwangsläufig gesetzt sind: die unermessliche Zahl unterschiedlicher Anwender, Umgebungen und Anforderungen lässt das Ziel einer umfassenden Methodenbibliothek utopisch erscheinen – einen Ausweg bieten in diesem Fall Anwenderprofile (vgl. Kapitel 4.3.3.1, vgl. King 2005:51).

Anwendungsszenarien

Der Evaluationsrahmen soll für anwenderorientierte Adäquatheitsevaluationen eingesetzt werden können und dabei einerseits Grundsätze der Softwareentwicklung nicht außer Acht lassen, anderseits Möglichkeiten bieten, wirklich die Bedürfnisse der potentiellen Anwender bzw. Käufer als Grundlage für die Evaluationskriterien zu verwenden. Evaluationsmetho-

dologien sind traditionell sehr auf die Bedürfnisse von Softwareentwicklern ausgerichtet. Mit dem EAGLES-Evaluationsrahmen sollen Grundlagen geschaffen werden, die sich vollständig der Anwendersicht widmen und die besonderen Bedürfnisse von sprachverarbeitender Software berücksichtigen (vgl. EAGLES 1995:67).

Das Ziel der Adäquatheitsevaluation ist zumeist ein Vergleich mehrerer Programme desselben oder eins ähnlichen Programmtyps – eine vergleichende Evaluation birgt besondere Schwierigkeiten: die Programme sind unterschiedlich aufgebaut, lösen einzelne Aufgaben auf verschiedene Weisen und bieten diverse oft nicht direkt vergleichbare Funktionen an – eine weitere Hürde besteht im fehlenden Bewusstsein der Anwender bzw. Käufer für ihre Anforderungen (vgl. EAGLES 1995:24). Das Ziel der Evaluation sollte in beiden Fällen nicht aus den Augen verloren werden, je genauer die Anforderungen definiert sind, desto klarer können einzelne Merkmale festgelegt werden. Zweckgebundenes Denken kann durchaus hilfreich sein (vgl. EAGLES 1995:53). Der Evaluator kann bei der Entwicklung von Anforderungen die Funktionen bestehender Produkte als Ausgangspunkt verwenden. Verschiedene Techniken zur Wissensakquisition (z.B. die KADS-Methode, vgl. EAGLES 1995:106ff.) bzw. bereits durchgeführte Evaluationen können ebenfalls hilfreiche Anhaltspunkte bieten. Darüber hinaus sollte daran gedacht werden, dass auch eine Nichterfüllung einer Anforderung ein sinnvolles Ergebnis darstellen kann (vgl. EAGLES 1995:103f.) und dass für den Anwender bzw. Käufer vor allem der informative Wert einer Evaluation zählt (Sparck-Jones & Galliers 1995:63).

4.3 Entwicklung von Anforderungen, Merkmalen & Maßen

In den nächsten Unterkapiteln werden die einzelnen Arbeitsschritte im Prozess der Definition von Merkmalen beschrieben. Zunächst wird auf die Analyse der Evaluationssituation eingegangen (Kap. 4.3.1), anschließend werden die Prinzipien für die Entwicklung von Anforderungen (Kap. 4.3.2) und die Definition von *wesentlichen informativen Merkmalen* (*reportable attributes*) mit den zugehörigen Maßen erläutert (Kap. 4.3.3). In Folge werden die grundlegenden Konzepte Gültigkeit und Zuverlässigkeit näher betrachtet (Kap. 4.3.4).

Für das ganze Kapitel – wie insgesamt für den EAGLES-Evaluationsrahmen – gilt, dass es sich um Vorschläge und unverbindliche Richtlinien handelt, die eine größere Vereinheitlichung und somit geteilte Nutzung von Ressourcen, Methoden und Ergebnissen ermöglichen sollen, jedoch keinesfalls um verpflichtende Regeln, die unumstößlich zu beachten sind (vgl. EAGLES 1995:73). Evaluationskonzepte müssen maßgeschneidert entwickelt werden, eine starre Allzweckdefinition ist nicht zielführend (vgl. Sparck-Jones & Galliers 1995: 193f.).

4.3.1 Analyse der Evaluationssituation

Bevor auf die Entwicklung der Qualitätsanforderungen eingegangen werden kann, müssen zahlreiche Faktoren analysiert werden. Zunächst muss klargestellt werden, für wen die Evaluation durchgeführt wird, welche Ziele verfolgt und welche weiteren Faktoren beachtet werden müssen.

4.3.1.1 Allgemeine vs. spezifische Evaluationen

Evaluationen, deren Design auf dem EAGLES-Evaluationsrahmen basiert, sollen dazu dienen, potentiellen Anwendern und Käufern von sprachverarbeitender Software Informationen zur Qualität und Eignung verschiedener Programme bereitzustellen. Grundsätzlich können zwei Formen der anwenderorientierten Adäquatheitsevaluation unterschieden werden: *spezifische Evaluationen* im Auftrag eines bestimmten Kunden und *allgemeine Evaluationen*, die verschiedenen potentiellen Anwendern und Käufern die Möglichkeit bieten, anhand der dargebotenen Information und eigenen Gewichtungen der Evaluationsergebnisse eine Entscheidung zu treffen (vgl. EAGLES 1995:41f.).

Der Arbeitsaufwand zur Feststellung der Anforderungen, anhand derer die Evaluationsmerkmale entwickelt werden sollen, kann in beiden Fällen sehr hoch sein: Für den Fall, dass es einen Auftraggeber der Evaluation gibt, sollten die Anforderungen so detailliert wie möglich an die Bedürfnisse des Auftraggebers angepasst werden. Manche Auftraggeber können ihre Bedürfnisse nicht ausreichend artikulieren, bzw. kennen weder die Möglichkeiten, die die Software bieten kann, noch die entsprechenden Voraussetzungen, unter denen sie sie besonders lukrativ nutzen könnten (vgl. EAGLES 1995:17).

Im anderen Fall, auf den sich die EAGLES-Projektgruppe überwiegend bezieht, steht bei der Durchführung der Evaluation noch nicht fest, welche potentiellen Anwender bzw. Käufer die Evaluation nutzen werden (vgl. EAGLES 1995:105). Diese Art der Evaluation orientiert sich am Consumer Report Paradigm, und ist vor allem durch Testberichte für Gebrauchsgüter bekannt (vgl. EAGLES 1995:13). In dieser Situation liegt es an den Evaluatoren mögliche Anwender- und Käuferprofile zu definieren, und Anforderungen für die verschiedenen Nutzungsmöglichkeiten aus unterschiedlichen Perspektiven zu entwickeln (vgl. EAGLES 1995:24). Zu einer Anpassung der Anforderungen gehört auch die Identifikation eines möglichen Informationsdefizits bzw. Erklärungsbedarfs, sowie die Notwendigkeit zusätzlicher Informationen zur Relevanz und möglichen Gewichtung festzustellen (vgl. EAGLES 1995:102).

Die im Rahmen dieser Arbeit entwickelten Ideen zur Methodologie sollen für beide Arten der anwenderorientierten Adäquatheitsevaluation genutzt werden können: es soll also Möglichkeiten geben, Anforderungen für potentielle Anwender und Käufer zu definieren, die

diese entweder selbst nicht identifizieren können bzw. die selbst noch nicht identifiziert worden sind.

Um die Anforderungen nichtsdestotrotz auf die zukünftigen Anwender bzw. Käufer abzustimmen, bietet sich die Anwendercharakterisierung (*user profiling*) an. Im Rahmen der Anwendercharakterisierung werden zahlreiche Charakteristika von relevanten potentiellen Anwendergruppen gesammelt, so dass Profile für einzelne Typen von Anwendern erstellt werden können, anhand derer in Folge die Anforderungen entwickelt werden können. Im Rahmen des EAGLES-Projektes wurde eine solche Übersicht an Anwenderprofilen für Hilfsmittel im Übersetzungsbereich entworfen, ausgehend von einer Beschreibung des Arbeitsbereiches, der Charakteristika der unterschiedlichen Aufgaben und Arbeitsumfelder sowie den Beziehungen zwischen diesen Charakteristika (vgl. EAGLES 1995:136ff. & 156-162).

4.3.1.2 Ziele, Setups & Umgebungsfaktoren

Im Vorfeld einer spezifischen Evaluation sollte feststehen, welches Ziel mit dem Erwerb und der Verwendung eines neuen Programmes verfolgt wird, und welcher Nutzen erwartet wird. Darüber hinaus muss klargestellt werden, wer das Programm benutzen wird und für welchen Zweck die jeweiligen Anwender es verwenden werden (Sparck-Jones & Galliers 1995:122). Im Vorfeld einer allgemeinen Evaluation sollte bereits feststehen, für welche potentiellen Anwendergruppen die Evaluation gedacht ist und welche Funktionen und Aufgaben im Rahmen der Evaluation abgedeckt werden sollen. Hierbei muss bereits auf die unterschiedlichen Anwender geachtet werden, die die Möglichkeit behalten müssen, nach der Evaluation selbst eine Entscheidung zu treffen, die nicht vorweggenommen werden soll (vgl. EAGLES 1995: 14).

Bei der Entwicklung von Anforderungen für spezifische potentielle Anwender bzw. Käufer, sollte das bereits bekannte zukünftige Setup (z.B. die vorhandene Hardware, die bereits installierten Programme) und die Umgebung (Zahl der Anwender sowie Kenntnisse der Anwender etc.) mit in die Definition der Anforderungen einbezogen werden. Auch im Rahmen von allgemeinen Evaluationen sollten die Einflüsse des Setups nicht außer Acht gelassen werden, sondern als potenzielle Faktoren Beachtung finden. Hierbei muss die mögliche Vielfalt in einer adäquaten Form abgebildet werden (vgl. EAGLES 1995:24f.).

4.3.2 Definition der Anforderungen

Anforderungen betreffen nicht nur Wünsche und positiv formulierte Erwartungen – es wird unterschieden zwischen *funktionalen Anforderungen* (*functional requirements*) und nicht-funktionalen Anforderungen. *Nichtfunktionale Anforderungen* (*non-functional requirements*) entsprechen Einschränkungen, Begrenzungen von Möglichkeiten, vorhandene bzw. bereits

erwartete Faktoren, die den Funktionsumfang einschränken können – z.B. die mangelnde Erfahrung des zukünftigen Anwenders im Umgang mit Computern (vgl. EAGLES 1995:23). Die zu erstellenden Anforderungen bestehen aus

- einer möglichst detaillierten, jedoch abstrakten Beschreibung der Aufgaben, die das zu evaluierende Programm im Rahmen einer potentiellen Anwendung bewältigen sollte,
- einer Beschreibung möglicher (bzw. des vorhandenen oder geplanten) Setups, in die das Programm integriert werden könnte (sollte),
- einer Beschreibung der relevanten Umgebungsfaktoren (anhand der Anwendercharakteristika) sowie
- einer detaillierten Beschreibung der funktionalen und nichtfunktionalen Anforderungen an jede einzelne Aufgabe (vgl. EAGLES 1995:26).

4.3.2.1 Erstellung eines Prozessmodells

Die erste Beschreibung der Aufgaben sollte möglichst abstrakt gehalten werden, so dass noch keine spezifische Umsetzung durch das Programm vorgeschrieben wird (vgl. EAGLES 1995: 110). Auf diese Weise soll verhindert werden, die Evaluation von Beginn an auf einen bestimmten Umsetzungstyp einzuschränken. Diese erste Beschreibung sollte sich möglichst nah an den Wünschen von potentiellen Anwendern orientieren, und für die Auftraggeber bzw. späteren Interessierten an der Evaluation nachvollziehbare und verständliche funktionale Vorgaben enthalten (vgl. EAGLES 1995:109f.). Dies kann z.B. anhand eines Prozessmodells mit den einzelnen Elementen des Programms, Rollen verschiedener Anwender und den Wegen von Datentypen im Verlauf der Anwendung des zu evaluierenden Programmtyps dargestellt werden (vgl. EAGLES 1995:109), z.B. für den Fall eines TEP:

> Ein **Terminologe** identifiziert **potentielle Termini** in einem **Textkorpus** und sortiert sie für einen **Experten**, der die **sortierten Termini** validiert und die **validierten Termini** in eine **Datenbank** exportiert.

4.3.2.2 Ausweitung des Prozessmodells

Im nächsten Schritt sollen das potentielle Setup (bzw. die potentiellen Setups) sowie die weitere Umgebung und hiermit verbundene Faktoren in die Struktur mit aufgenommen werden. Das Prozessmodell kann also ausgeweitet werden, für Unterteilungen können unterschiedliche Ebenen verwendet werden, die Beschreibungen der einzelnen Aufgaben vererbt bzw. modifiziert werden (vgl. EAGLES 1995:113). Die Eingabe und Ausgabe von Daten, bereits vorhandene Programme, die mit dem neuen Programm kommunizieren sollen und Doku-

mentformate werden hier in das Modell eingefügt (vgl. EAGLES 1995:27). Zwei Beispiele sollen die Spezifizierung verdeutlichen:

> Ein **Assistent** stellt einen **Textkorpus** aus verschiedenen **HTML- und PDF-Dokumenten** zusammen, identifiziert **potentielle Termini** und sortiert sie anhand von **Zusatzinformationen** nach Fachgebieten. Verschiedenen **Experten** werden die ihnen zugeordneten **sortierten Termini** in einem betriebsinternen **CMS** angezeigt, die **Experten** validieren die **sortierten Termini** anhand der **Zusatzinformationen** und der Assistent exportiert die **validierten Termini** in eine **Multiterm-Datenbank**.
>
> Ein **Terminologe** stellt einen **Textkorpus** aus verschiedenen **Microsoft Word-Dokumenten** zusammen und identifiziert **potentielle Termini**, die er mit bereits **vorhandener Terminologie** aus einer **Datenbank** abgleicht. Ein **Experte** validiert die **neuen potentiellen Termini** und exportiert die **validierten Termini** in eine **Datenbank**. Beim Verfassen **neuer Texte** durch **Redakteure** wird die **Verwendung der Terminologie** aus der **Datenbank** überprüft.

4.3.2.3 Formulierung der Anforderungen

Im letzten Schritt werden explizite funktionale und nichtfunktionalen Anforderungen in das Modell integriert, so dass eine komplexe Struktur mit übergeordneten Aufgaben, relevanten Einflussfaktoren und funktionalen sowie nichtfunktionalen Anforderungen die Datenströme ins und aus dem Programm über verschiedene Aufgaben unter der Einwirkung unterschiedlicher Anwender repräsentiert (vgl. EAGLES 1995:109). Die an dieser Stelle hinzugefügten Anforderungen sollen sich auf die übergeordneten Aufgaben beziehen und keine Vorgaben für die Umsetzung der Aufgabe in softwaretechnischer Hinsicht beinhalten (vgl. EAGLES 1995:23 & 28). Eine Modularisierung der Anforderungen nach einzelnen Komponenten bzw. Aufgabenbereichen bietet sich zum Zwecke der Übersichtlichkeit an, auch wenn die Trennung der Auswirkungen einzelner Komponenten im Rahmen einer Black-Box-Evaluation nur eingeschränkt möglich ist (vgl. EAGLES 1995:24). Als Kategorien für eine Modularisierung können auch die einzelnen Kriterien des in Kapitel 4.1.2 vorgestellten Qualitätsmodells verwendet werden. Auf diese Weise können auch bereits jene im Modell enthaltenen Qualitätskriterien ausgewählt werden, die im Rahmen der Evaluation prioritär behandelt werden sollen, z.B. Funktionalität, Benutzbarkeit, Zuverlässigkeit und Anpassbarkeit (vgl. EAGLES 1995:121f.). Anforderungen könnten beispielsweise wie folgt lauten:

- Das Programm sollte im Textkorpus enthaltene relevante Termini als potentielle Termini identifizieren.

- Das Programm sollte Möglichkeiten bieten, die Vorgaben für relevante Termini zu spezifizieren.
- Das Programm sollte große Textkorpora im Umfang von x Normseiten Text bearbeiten können.[26]

Um die auf diesem Wege entwickelten Anforderungen im Rahmen der Evaluation zu verwenden, müssen sie zunächst in sogenannte *reportable attributes* (die in dieser Arbeit als *wesentliche informative Merkmale* bezeichnet werden sollen) umgewandelt werden (vgl. EAGLES 1995:103). Dieser Schritt soll im nächsten Kapitel betrachtet werden.

4.3.3 Merkmale & Maße

Sobald die Anforderungen an die Evaluation und an die zu evaluierenden Programm feststehen, können *wesentliche informative Merkmale* (*reportable attributes*) entwickelt werden: wesentliche informative Merkmale sind Merkmale, dessen Überprüfung im Rahmen der Evaluation zu *Werten* (*values*) führt, die für potentielle Anwender und Käufer interessant und hilfreich bei der Entscheidung über die Adäquatheit eines Programms sein können (vgl. EAGLES 1995:29). Durch die Summe aller definierten Merkmale sollten möglichst sämtliche explizite und implizite Anforderungen der Anwender abgedeckt werden – gleichzeitig sollten Merkmale allgemein genug formuliert sein, um auf unterschiedliche Programme und für unterschiedliche Anwendergruppen verwendet werden zu können (vgl. EAGLES 1995: 21).

Für die Definition der wesentlichen informativen Merkmale müssen in manchen Fällen bestehende Programme und ihre Funktionen und Schwächen analysiert werden, um herauszufinden, welche Merkmale von Programmen des entsprechenden Programmtyps entscheidungsrelevante Informationen liefern könnten – nichtsdestotrotz sollte die Umsetzungsform von einzelnen Funktionen nicht vorgeschrieben werden, um keine Programme von vornherein zu benachteiligen. Hierfür müssen unter Umständen alle bekannten Umsetzungswege angeführt werden, sowie Möglichkeiten geboten werden, neue bisher unbekannte Umsetzungswege gleichberechtigt zu dokumentieren (vgl. EAGLES 1995:85 & 112).

4.3.3.1 Rolle der Anwenderprofile

An dieser Stelle müssen auch die Anwenderprofile bedacht werden. Zukünftige Anwender finden eventuell andere Merkmale relevant als potentielle Käufer (vgl. Sparck-Jones & Galliers 1995:39). Auch innerhalb einzelner Anwendergruppen können unterschiedliche Kriterien von Bedeutung sein, bzw. dieselben Kriterien unterschiedlich ausgelegt werden. Dies soll am

[26] Die Originalbeispiele, die bei der Entwicklung dieser Beispiele als Grundlage dienten, beziehen sich auf Grammatikprüfprogramme (vgl. EAGLES 1995:127).

Beispiel der Terminologieextraktion verdeutlicht werden: wie zu Beginn der Arbeit in Kap. 1.1 festgestellt erwartet jeder Anwender eines TEP ein anderes Extraktionsergebnis, da jeder eigene Vorstellungen davon hat, wie relevante Termini zu definieren sind. Dies muss bei der Umwandlung von Anforderungen in Merkmale bedacht werden, um nicht einzelne Anwendergruppen auszuschließen. Eine Möglichkeit wäre, eine Anpassung des Merkmals, z.B. durch die Definition eines kombinierten Merkmals aus Precision und Recall-Werten, mit bestehenden Modifizierungsmöglichkeiten zuzulassen. Eine andere Möglichkeit besteht darin, zahlreiche voneinander getrennte Merkmale anzugeben, die der Interessent als Zusammenspiel mehrerer Faktoren selbst gewichten kann (vgl. Sparck-Jones & Galliers 1995:21f.). Das Problem der divergierenden Vorstellung davon, was ein Terminus ist, kann jedoch auch durch diese Varianten nicht gelöst werden.

Falls eine allgemeine Evaluation angestrebt wird, können die Anforderungen unterschiedlicher Anwendergruppen zu kontradiktorischen Merkmalen führen. In diesem Fall liegt es am späteren Anwender der Evaluation, der für sich entscheiden sollte, welche Merkmale seinen Anforderungen entsprechen (vgl. EAGLES 1995:29). Da die Merkmale im Bereich der Softwareevaluation eventuell nicht allen Interessenten bekannt sind, muss bei der Dokumentation der Evaluation daran gedacht werden, potentielle Unklarheiten zu vermeiden und erklärende Informationen hinzuzufügen, um den Interessenten eine Auswahl der für sie relevanten Merkmale zu erleichtern (EAGLES 1995:130f.). Der Evaluator sollte sich für seine Auswahl an wesentlichen informativen Merkmalen daran orientieren, welche Merkmale für besonders viele potentielle Interessenten relevant und nützlich als Entscheidungshilfe sein könnten (vgl. EAGLES 1995:131). Auch in diesem Fall gilt, der informative Wert als Ziel der Evaluation sollte nicht aus den Augen verloren werden (vgl. Sparck-Jones & Galliers 1995:63).

4.3.3.2 Zuordnung von Maßen

Jedem Merkmal muss eine entsprechende Vorgabe an möglichen Werten zugeordnet werden. Dies soll an einem Beispiel verdeutlicht werden: z.B. kann das Merkmal lauten *Möglichkeit des Exports in eine Datenbank*. In diesem Fall können verschiedene Wertvorgaben getätigt werden: Entweder soll der Wert lediglich aussagen, ob die Möglichkeit des Exports besteht oder nicht; oder es soll spezifiziert werden, in welchen Datenformaten Daten exportiert werden können. Merkmale werden entsprechend der Art von Werten, die ihnen zugordnet werden, in verschiedene Typen eingeteilt werden: Binäre (z.B. ja/nein), klassifizierende (z.B. linguistisch/statistisch/hybrid), vergleichende (z.B. eine Bewertungsskala von sehr einfach bis sehr schwierig), numerische (z.B. eine beliebige Zahl) oder metrische Werte (z.B. ein Prozentsatz) (vgl. EAGLES 1995:31).

Der erwartete Wertetyp einschließlich der möglichen Werte wird als *Maß* (*measure*) bezeichnet (vgl. EAGLES 1995:29) und muss im Rahmen der Definition des Merkmals festgelegt werden. Ein mit einem Wertetyp und vorgegebenen möglichen Werten versehenes Merkmal ergibt ein *Kriterium* (*feature*). Der Evaluationsrahmen fordert für Merkmale, dass die zu erhaltenden Werte durch Beobachtung, direkte Messung oder Ableitung von den Werten anderer Merkmale festgestellt werden können (vgl. EAGLES 1995:20).

4.3.4 Gültigkeit & Zuverlässigkeit

Für alle Ebenen der Anforderungen, Merkmale und Maße gelten die Grundsätze der Gültigkeit (*validity*) und der Zuverlässigkeit (*reliability*). Jeder einzelne Schritt muss gültig und zuverlässig sein, sonst ist die Zuverlässigkeit und Gültigkeit des Evaluationsergebnisses eingeschränkt. Dieser Prozess beginnt bereits bei der Analyse der Anforderungen (vgl. EAGLES 1995:28) und ist besonders bei der Entwicklung von Maßen von Bedeutung (vgl. EAGLES 1995:29f.).

Die Gültigkeit von Maßen kann unterteilt werden in interne und externe Gültigkeit: die interne Gültigkeit besteht, wenn ein Maß tatsächlich die Interessen der Anwender bzw. Käufer wiederspiegelt und misst. Externe Gültigkeit wird durch eine vermutete oder bestätigte Korrelation mit einem gültigen Maß erreicht – der Bezug zu den Anwenderinteressen ist in diesem Fall nur indirekt vorhanden (vgl. EAGLES 1995:30). Die Zuverlässigkeit von Maßen bezieht sich auf die Möglichkeit durch eine Methode beständige Ergebnisse des jeweiligen Maßes zu erhalten, ohne situationsbedingte Schwankungen, die die Zuverlässigkeit des Maßes einschränken würden. Um eine hohe Zuverlässigkeit zu erreichen, sollte möglichst auf menschliche Beteiligung und Interaktion sowie auf subjektive Beurteilungen bei der Durchführung der Evaluation verzichtet werden (vgl. EAGLES 1995:30f.).

4.4 Methodologie

Im Rahmen einer Evaluation können sehr unterschiedliche Aspekte betrachtet und evaluiert werden. Abhängig von der Situation, den beteiligten Personen, den zu evaluierenden Programmen, dem vorhandenen Budget und Material sowie der zur Verfügung stehenden Zeit wird für jede Evaluation die entsprechende Methodologie festgelegt. Die Methodologie für eine Evaluation besteht aus der Auswahl eines oder mehrerer Testtypen und der Ausarbeitung und Kombination der einzelnen Methoden (anhand von Anforderungen und Merkmalen) einschließlich des zu verwendenden Materials sowie der Festlegung der zugehörigen Dokumentation. Diese Entscheidungen müssen für jede Evaluation erneut anhand der jeweiligen Situation und des Evaluationsziels getroffen werden (EAGLES 1995:73).

4.4.1 Art der Evaluation

In dieser Arbeit sollen keine Evaluationen von Programmen im Entwicklungsstadium betrachtet werden, sondern der Fokus auf verfügbare, eventuell noch anpassbare, in jedem Fall bereits vollständige und funktionsfähige Programme gerichtet werden. Es kann bei diesen Programmen nicht davon ausgegangen werden, Einblick in die innere Struktur erhalten und einzelne Komponenten im Rahmen von Glass-Box-Evaluationen untersuchen zu können. Aus diesem Grund werden lediglich Black-Box-Evaluationsmethoden betrachtet.

4.4.1.1 Anwenderorientiert

Darüber hinaus konzentriert sich diese Arbeit auf die anwenderorientierten Evaluation, die zum Zeitpunkt des EAGLES-Projektes in der existierenden Literatur in geringerem Maße betrachtet wurde, der stärkere Fokus lag auf der entwicklerorientierten Evaluation (vgl. Sparck-Jones & Galliers 1995:146). Anwenderorientierte Evaluationen verlangen eine angepasste Methodologie, da andere Informationen erhoben werden müssen: entwicklerorientierte Ansätze können dementsprechend nicht unverändert übernommen werden (vgl. EAGLES 1995:67). An dieser Stelle soll erneut darauf hingewiesen werden, dass anwenderorientiert nicht bedeutet, dass Anwender an der Evaluation beteiligt sein müssen. Die Beteiligung von Anwendern an einer Evaluation ist unabhängig von ihrem anwenderorientierten oder entwicklerorientierten Fokus (vgl. EAGLES 1995:71).

4.4.1.2 Systemvergleich

Für viele denkbare Szenarien des EAGLES-Evaluationsrahmens spielen Vergleiche verschiedener Programme eine wichtige Rolle, das Interesse an vergleichenden Evaluationen spiegelte sich bereits in den 90ern auch in der Literatur wider (EAGLES 1995:67). Für Merkmale, denen komparative Werte zugeordnet werden, können vergleichende Evaluationen die einzige Möglichkeit darstellen, nützliche Ergebnisse zu erhalten – es ist jedoch schwierig, im Rahmen von vergleichenden Evaluationen zuverlässige Methoden zu finden (vgl. EAGLES 1995: 149).

Darüber hinaus muss bei der Definition der Methoden und der Festlegung bzw. Vorgabe von Charakteristika des Testmaterials darauf geachtet werden, dass die getesteten Programme unterschiedliche Ansätze verfolgen können, und somit durch eingeschränkt repräsentatives Material oder bestimmte Vorgaben eventuell einzelne Programme benachteiligt werden könnten (vgl. Sparck-Jones & Galliers 1995:58). Für vergleichende Evaluationen empfiehlt es sich, besonders darauf zu achten eine möglichst umfassende Auswahl an Programmen zu treffen, bereits im Vorfeld möglichst objektive und standardisierte Maße fest-

zulegen, essentielle Unterschiede zu betonen und sowohl positive als auch negative Ergebnisse möglichst strukturiert und übersichtlich wiederzugeben (vgl. EAGLES 1995:46f.).

4.4.2 Evaluationsdesign

Grundsätzlich können zwei unterschiedliche Ansätze verfolgt werden, ein experimenteller oder ein investigativer Ansatz. Für möglichst exakte, objektive Ergebnisse und die Überprüfung der Auswirkung von bestimmten Faktoren sollte ein Experiment durchgeführt werden. Im Rahmen eines Experiments müssen alle möglichen Variablen kontrolliert werden – hierfür ist eine aufwendige Vorbereitung und Validierung des Materials und der Bedingungen notwendig. Falls dieser Aufwand nicht möglich ist, oder auf flexiblere, ergebnisoffenere Weise Informationen gesammelt werden sollen, kann der investigative Ansatz gewählt werden. Unabhängig davon, dass die Vorbereitung unterschiedlich aufwendig ausfallen kann, muss für beide Ansätze eine entsprechende Methodologie ausgewählt oder entwickelt werden (vgl. Sparck-Jones & Galliers 1995:22).

Der gesamte Prozess von der Betrachtung der zu evaluierenden Programme bis hin zur Definition von wesentlichen informativen Merkmalen kann langwierig und aufwendig sein – das EAGLES-Projekt strebt die Sammlung, Wiederverwendung und Weiterentwicklung von Anwendercharakteristika, Anforderungen, Merkmalen, Maßen und Methoden an und sieht den Prozess mit dem Ende einer Evaluation nicht als abgeschlossen an. Zunächst muss jedoch in jedem einzelnen Gebiet Pionierarbeit geleistet werden, um erste Ergebnisse zu erzielen, auf denen in weiterer Folge aufgebaut werden kann (vgl. EAGLES 1995:112).

In den folgenden Unterkapiteln sollen zunächst Methoden und ihre Eigenschaften und in Folge die einzelnen Testtypen, Testmaterialien und Dokumentationsmöglichkeiten vorgestellt werden. Im Anschluss daran wird in einer Übersicht der gesamte Ablauf zur Vorbereitung und Durchführung einer Evaluation skizziert (siehe Kap. 4.5).

4.4.3 Methoden

Im Rahmen dieses Unterkapitels sollen wichtige Eigenschaften von Methoden (einschließlich des Testmaterials und der Vorgaben zur Interpretation) hervorgehoben werden, die vor und bei der Planung einer Evaluation bedacht werden sollten.

Der Hauptbestandteil jeder Methodologie sind die Methoden: Für jedes definierte Kriterium (Merkmal-Maß-Paar) muss eine zuverlässige und gültige Methode gefunden werden, die dazu dient, die einzelnen Werte der jeweiligen Merkmale zu bestimmen. Auf jene Kriterien, für die keine gültige und zuverlässige Methode gefunden werden kann, sollte im Zweifelsfall lieber verzichtet werden (EAGLES 1995:171). Darüber hinaus sollten Methoden effizient anwendbar, möglichst formalisiert und nützlich für das Evaluationsziel sein (vgl.

EAGLES 1995:148). Die Beschreibung einer Methode enthält neben einer Beschreibung der durchzuführenden Beobachtung, Messung oder sonstigen Tätigkeit auch Informationen über die zu verwendenden Materialien sowie über die Vorgaben zur Interpretation des Ergebnisses (vgl. EAGLES 1995:21 & 132). Die Methode soll nur den Wert des jeweiligen Merkmals bestimmen, andere Merkmale sollten so wenig Einfluss wie möglich auf das Ergebnis haben. Falls die Werte einer kombinierten Methode bestimmt werden sollen, muss die Gewichtung der einzelnen Bestandteile im Vorfeld transparent festgelegt werden (vgl. EAGLES 1995: 133f.). In Folge sollen weitere wichtige Bestandteile von Methoden näher betrachtet werden.

4.4.3.1 Testmaterial: Umfang, Realitätsnähe & Repräsentativität

Sprachverarbeitende Programme verarbeiten Sprache in verschiedenen Formen, oftmals als geschriebenen, elektronisch lesbaren Text. Zum Testen wird ebenfalls Material benötigt, und es ist wichtig, sich bewusst für die Art des Materials und der Erfassung zu entscheiden, da das Material starke Auswirkungen auf Testergebnisse haben kann (vgl. L'Homme 1996:303).

Testmaterial kann auf zwei Faktoren überprüft werden: Realitätsnähe und Repräsentativität – die Realitätsnähe bezeichnet den Grad, in dem das Testmaterial realen Kontexten entnommen wurde und auch in einer realen Situation für eine derartige Anwendung verwendet werden könnte. Die Repräsentativität bezieht sich auf Texttypen und Textelemente des Testmaterials, deren Verteilung und Häufung jeweils den Rahmen der Materialien und Bereiche vorgibt, für den das Material als repräsentativ angesehen werden kann (vgl. Sparck-Jones & Galliers 1995:148). Für die Zuverlässigkeit der Methode ist es wichtig, ein Minimum an zu verwendenden Testmaterialien festzulegen (vgl. EAGLES 1995:132).

Für den Fall, dass die zu evaluierenden Programme auf den Umgang mit bestimmten Phänomenen getestet werden sollen bzw. bei Programmen desselben Programmtyps Probleme im Umgang mit bestimmten Phänomenen bekannt sind, kann das Testmaterial besonders ausgewählt bzw. bearbeitet werden (vgl. Kap. 4.4.6), um möglichst viele dieser Phänomene zu enthalten. Es sollte jedoch nicht versucht werden, im Rahmen einer Evaluation sämtliche möglicherweise auftretenden Phänomene durch das Testmaterial abzudecken: das ist einerseits utopisch und anderseits nicht zielführend: eine weniger umfassend gesteckte Evaluation, bei der größter Wert auf Gültigkeit und Zuverlässigkeit gelegt wurde, kann sich als nützlicher erweisen als eine Evaluation, die auf derart umfassendem Testmaterial basiert, dass der Arbeitsaufwand nicht mehr bewältigbar ist und ihre Zuverlässigkeit und Gültigkeit eventuell nicht garantiert werden können. Vor allem im Bereich der sprachverarbeitenden Programme müssen Testergebnisse oft von Menschen beurteilt werden, da eine Automatisierung der Validierung nicht möglich ist – insbesondere in solchen Fällen muss darauf geachtet werden, dass das Testmaterial eine bewältigbare Größe behält (vgl. EAGLES 1995:133).

4.4.3.2 Interpretation der Ergebnisse: Vorgaben, Baselines & Benchmarks

Im Vorfeld der Evaluation müssen bereits Vorgaben für die spätere Interpretation der Ergebnisse definiert werden. Dadurch soll verhindert werden, dass im Nachhinein willkürliche Beurteilungen getroffen werden. Solche Vorgaben können sich an den Anforderungen orientieren und z.B. Mindeststandards festlegen. Zur Veranschaulichung einige Beispiele:

- *Der Noise-Anteil eines TEP ist als zu hoch zu bezeichnen, falls mehr als 50% des Identifikationsergebnisses im Zuge der Validierung als inkorrekt bezeichnet wird, bei weniger als 25% wird der Noise-Anteil als gering bezeichnet.*
- *Die Nachbearbeitungsfunktion wird folgendermaßen klassifiziert:*
  Erfüllung von 7-10 der gewünschten Funktionen – umfassend, 5-7 ausreichend, 0-5 nicht ausreichend.

Bei den Vorgaben zur Interpretation der Ergebnisse sollten einige wichtige Aspekte beachtet werden. Beim Betrachten und Vergleichen von Ergebnissen, besonders jenen, die die Leistung beschreiben, sollte darauf geachtet werden, die Umgebungsfaktoren mit einzubeziehen, und somit eine gerechte Interpretation der Ergebnisse durchzuführen (vgl. Sparck-Jones & Galliers 1995:45).

Grundsätzlich ist es schwer, direkte Bestleistungsziele in Zahlen oder Ergebnissen vorzugeben, die erreicht werden sollen/müssen, und es sollte in jedem Fall darauf verzichtet werden perfekte und somit irreale Ergebnisse anzustreben (vgl. Sparck-Jones & Galliers 1995: 148). Im Bereich des Information Retrieval hat sich gezeigt, dass Bestleistungsziele aufgrund der mangelnden Eindeutigkeit kaum formulierbar sind. Eine auf einer anerkannten *baseline*[27] – eine grundlegende Leistung eines Vergleichssystems – basierende Evaluation, die dadurch realistisch erreichbare Ziele enthält, bzw. eine vergleichende Evaluation mehrerer Systeme sind geeigneter und liefern informativere Ergebnisse. Baselines sind in gewissem Sinne minimale Vorgaben: sie können neben der Beschreibung eines grundlegenden Funktionsumfangs z.B. einen Durchschnittswert der Leistung von bereits evaluierten Programmen enthalten. Leider gibt es in zahlreichen Bereichen der sprachverarbeitenden Software noch keine definierten Baselines (vgl. Sparck-Jones & Galliers 1995:139).

Eine weitere Möglichkeit besteht in der Betrachtung quantitativer Ergebnisse im Vergleich zu *benchmarks*. Benchmarks sind die jeweils erreichten Höchstwerte oder in einigen Fällen die zu erreichenden Ziele. Im Rahmen eines Tests können zahlreiche unterschiedliche Benchmarks angewendet werden, eine Benchmark kann zum Beispiel aus an bestimmtes

[27] Unter Baseline wird in dieser Arbeit das Ergebnis der Anwendung eines Programms mit einer ähnlichen, aber nicht identischen Funktion gesehen, dass als Ausgangsbasis und Mindestanforderung genutzt werden kann (vgl. z.B. Witschel 2005:72).

Testmaterial angepassten Vorgaben für das Ergebnis bestehen – diese müssen aber auch Umgebungsfaktoren miteinschließen und sollten kein perfektes Ergebnis verlangen (vgl. Sparck-Jones & Galliers 1995:139).

Evaluationsergebnisse sollten nicht direkt mit den Ergebnissen von Menschen, die die Aufgaben manuell absolviert haben, verglichen werden. Es gibt keinerlei Garantie dafür, dass die menschlichen Ergebnisse perfekt sind – auch Menschen unterlaufen Fehler. Vermutlich sind die Ergebnisse zusätzlich subjektiv beeinflusst (vgl. Sparck-Jones & Galliers 1995:148 & EAGLES 1995:110). Die Verwendung von Computerprogrammen kann darüber hinaus Vorteile haben, die in einem direkten Vergleich selten beachtet werden, z.B. die Dokumentation der Arbeit, die Genauigkeit oder Geschwindigkeit (vgl. Ahmad & Rogers 2001:740).

Laut Sparck-Jones & Galliers sollte eine Evaluation an der zu erfüllenden Aufgabe ausgerichtet werden, nicht an einer Idealleistung (vgl. 1995:144).

4.4.3.3 Eignung: Anpassung & Granularität

Sprachverarbeitende Programme entsprechen nur selten genau den Anforderungen, die ein potentieller Anwender bzw. Käufer an ein Programm hat. In zahlreichen Fällen müssen die Programme vor der Anwendung in unterschiedlichem Maße an den zukünftigen Anwender angepasst werden. Dies kann sowohl interne Programmkomponenten betreffen als auch die Anpassung von Parametern bzw. die Bereitstellung von Test- bzw. Trainingsmaterial. Eine solche Anpassung findet selbst im Rahmen einer spezifischen Evaluation üblicherweise nicht statt (vgl. EAGLES 1995:51).

Falls doch Parameter angepasst und Faktoren verändert werden, spielt die Granularität – die Detailliertheit des Rasters, in dem Änderungen vorgenommen werden – eine entscheidende Rolle: werden sehr kleine Änderungen vollzogen, um die Auswirkungen von Details auf das jeweilige Programm zu testen, oder im Zuge jeder Anpassung der Faktoren grundlegende Aspekte geändert. Die Granularität kann Auswirkungen auf das Evaluationsergebnis haben und sollte im Rahmen der Dokumentation beschrieben werden, um die Vergleichbarkeit zu anderen Studien zu erleichtern (vgl. Sparck-Jones & Galliers 1995:26).

Darüber hinaus sollte bei der Beschreibung von Methoden bereits darauf geachtet werden, dass deutlich wird, welche spezifischen Merkmale beschrieben werden, und ob und in welcher Art Parameter eingestellt bzw. Umgebungsfaktoren beeinflusst wurden, so dass keine ungerechtfertigten Generalisierungen bei der Interpretation der Ergebnisse erfolgen (vgl. Sparck-Jones & Galliers 1995:26f.).

4.4.4 Testtypen

Den Rahmen für angewendete Testmethoden, Materialien und Instrumente bietet ein Testverfahren (oder eine Kombination aus Testverfahren). Es gibt zahlreiche Testtypen, die unterschiedliche Rahmenbedingungen für die Testphase bieten. Hauptsächlich beziehen sich Klassifikationen in diesem Bereich auf entwicklerorientierte Evaluationen (vgl. Hoffmann 2008:159 & 171). In diesem Kapitel soll lediglich eine Auswahl der Testtypen behandelt werden, die im EAGLES-Bericht diskutiert werden. Jene Testtypen, die sich nur für Glass-Box-Verfahren eignen, werden nicht beachtet.

Es werden folgende Testtypen im Rahmen dieses Kapitels vorgestellt: *Szenario-Tests, systematische Tests* und die *Kriterienüberprüfung* (vgl. EAGLES 1995:32f.). Szenario-Tests werden unter der Beteiligung von Anwendern durchgeführt, an systematischen Tests und Kriterienüberprüfungen sind keine Anwender beteiligt (vgl. EAGLES 1995:33ff.).

4.4.4.1 Szenario-Tests

In Szenario-Tests werden möglichst authentische Arbeitsabläufe getestet, um das zu testende Programm in alltäglichen Arbeitssituationen zu überprüfen. Dies kann entweder in der natürlichen Umgebung des Anwenders geschehen (*Feldtest*) oder aber in einer speziell eingerichteten Prüfumgebung (*Labortest*). Es wird eine standardisierte Aufgabe vorgegeben, die die Anwender mit dem jeweiligen Programm durchführen müssen. Die Planung und Vorbereitung von Szenariotests ist aufwendig, von großer Bedeutung ist die Dokumentation, die die Datenanalyse und Beurteilung im Nachhinein ermöglicht. Die Objektivität der Ergebnisse von Szenario-Tests ist eingeschränkt, jedoch können sie sehr informativ bei den Evaluationen von Kriterien der Benutzbarkeit (z.B. Verständlichkeit, Erlernbarkeit, Bedienbarkeit) sein (vgl. EAGLES 1995:74-80).

4.4.4.2 Systematische Tests

Systematische Tests werden von Evaluatoren unter spezifischen Bedingungen durchgeführt: es gibt Vorgaben für durchzuführende Aufgaben und je nach Unterart werden bestimmte Ergebnisse erwarte, z.B. kann die Funktionalität oder Laufzeit einer bestimmten Funktion getestet werden. Systematische Tests sind für Adäquatheitsevaluationen besonders gut geeignet, aus diesem Grund sollen sie hier etwas näher betrachtet werden. Im EAGLES-Projekt wurden drei Unterarten des systematischen Tests identifiziert: *Aufgabenorientiertes*, *menüorientiertes* und *Benchmark-Testen* (vgl. EAGLES 1995:34).

In einem aufgabenorientierten Test wird ebenfalls eine bestimmte Funktion getestet. Im Vergleich zum Szenario-Test ist dieser Test mit weniger Aufwand verbunden: der Task wird definiert, die Testmethoden vorbereitet und direkt von Evaluatoren durchgeführt, dokumen-

tiert und beurteilt. In aufgabenorientierten Tests können vor allem Informationen zur Funktionalität (einschließlich der Angemessenheit, Richtigkeit und Interoperabilität) erhoben werden und überprüft werden, ob ein Programm wirklich eine Aufgabe oder Anforderung erfüllt. Die Ergebnisse können aufgrund der schlechten Vergleichbarkeit unterschiedlicher Programme jedoch subjektiv beeinflusst sein (vgl. EAGLES 1995:34 & 81f.).

Im menüorientierten Test wird systematisch jede einzelne Funktion des Programms aufgerufen. Im Vorfeld wird keine bestimmte Aufgabe bestimmt, die zu sammelnden Daten und Testmethoden werden *ad hoc* bestimmt. Dadurch ist es für einen menüorientierten Test wichtig, über erfahrene Evaluatoren zu verfügen. Es können sehr unterschiedliche Ergebnisse entstehen, die nicht vorausgesehen werden können. Derartige Tests sind geeignet, um die gesamte Funktionspalette zu entdecken, keine Funktion zu übersehen und die Evaluation so adaptiv wie möglich zu gestalten sowie um die Zuverlässigkeit (einschließlich der Reife, Fehlertoleranz und Wiederherstellbarkeit) von Programmen zu überprüfen (vgl. EAGLES 1995:34 & 82ff.).

Das Ziel eines Benchmarktests ist eine möglichst objektive Überprüfung der Effizienz und Laufzeit von Programmen für die Durchführung bestimmter Funktionen bzw. Aufgaben. Ein Benchmarktest wird aufwendig vorbereitet, es sollten möglichst viele Faktoren kontrolliert werden (einschließlich der Eingabedaten), um objektive metrische und somit vergleichbare Ergebnisse zu erhalten. Ein Benchmarktest kann unter Umständen automatisiert werden und dient vor allem, um minimale Anforderungen zu überprüfen bzw. die Leistungsfähigkeit zu vergleichen und zur Evaluation der Effizienz (einschließlich des Zeit- und Verbrauchsverhaltens) von Programmen (vgl. EAGLES 1995:34 & 84).

4.4.4.3 Kriterienüberprüfung

Eine weitere Art der Evaluation, die nicht direkt den Testtechniken zugeordnet werden kann, ist die Kriterienüberprüfung (*feature inspection*). Ziel der Kriterienüberprüfung ist es, eine möglichst detaillierte Untersuchung der Programme durchzuführen. Der Vorbereitungsaufwand ist entsprechend groß, da eine detaillierte Kriterienchecklisteerstellt werden muss, die die Kriterien und die Methoden zur Überprüfung enthalten muss. Diese Checkliste baut auf dem Wissen über bekannte Programme, ihren Funktionen und Eigenschaften sowie Problemen auf. Mithilfe einer Kriterienüberprüfung kann die Funktionalität von Programmen evaluiert werden. Durch eine Konzentration auf binäre bzw. klassifizierende Werte kann ein hoher Grad an Objektivität erreicht werden (vgl. EAGLES 1995:84f.).

4.4.4.4 Validierung von Tests

Tests sollten im Anschluss an die Durchführung validiert werden: Für die Validierung von Tests wird überprüft, ob die verwendeten Maße tatsächlich Einfluss auf Nützlichkeit eines Programmes haben. Dies geschieht zum Beispiel durch eine Messung der Produktivität und/oder Zufriedenheit realer Anwender des jeweiligen Programmes. Da ein Programm tatsächlich unter realen Bedingungen in Verwendung sein muss, um eine Validierung durchführen zu können, gestaltet sich der Validierungsprozess aufwendig und entsprechend zeit- und kostenintensiv. Manche Maße werden durch die oben genannte interne Gültigkeit als gültig angesehen, z.B. Geschwindigkeit der Anwender beim Durchführen von Aufgaben – es ist allerdings unklar, ob die Produktivität und Zufriedenheit von Anwendern überhaupt durch die Verwendung bestimmter Programme gesteigert werden können (vgl. EAGLES 1995:150).

4.4.4.5 Testtypen & Qualitätskriterien

An dieser Stelle kann es hilfreich sein, sich zu vergegenwärtigen, welche Testtypen für welche Kriterien des Qualitätsmodells der ISO-IEC Norm 9126 besonders geeignet sind. Im oberen Kapitel wurden folgende Eignungen für die Kriterien der internen und externen Qualität festgestellt:

- Funktionalität → *Aufgabenorientierter systematischer Test, Kriterienüberprüfung*
- Zuverlässigkeit → *Menüorientierter systematischer Test*
- Benutzbarkeit → *Szenario-Test*
- Effizienz → *Benchmarktest*
- Änderbarkeit
- Übertragbarkeit

Die ersten vier Kriterien, denen besonders passende Testtypen zugeordnet wurden, werden von Hoffmann als für Anwender sichtbare Qualitätskriterien bezeichnet (vgl. 2008:9). Die bisher nicht zugeordneten Kriterien Änderbarkeit und Übertragbarkeit gehören zu den inneren Werten der Software, die für Anwender ohne erweiterte Kenntnisse schwerer erkennbar, nichtsdestotrotz vor allem in langfristiger Hinsicht entscheidungsrelevant sein können (Hoffmann 2008:10). In einigen Fällen ist es schwierig, alle Teilkriterien abzudecken, so erfordern z.B. zwei Teilkriterien der Funktionalität – Angemessenheit und Richtigkeit – besondere Aufmerksamkeit. Ein Programm kann richtige Ergebnisse im Sinne des Programms liefern, die für die Zwecke des Anwenders alles andere als angemessen sind. Eine Verbesserung der Angemessenheit ist in den meisten Fällen von der Unterstützung durch den Anwender abhängig, da seine Bedürfnisse bei der Anwendung des Programmes in Betracht gezogen werden müssen, um für ihn angemessene Ergebnisse zu produzieren. Sobald die Interaktivität mit

dem Anwender bedeutende Auswirkungen hat, erschwert dies die Möglichkeiten, eine objektive Evaluation durchzuführen (vgl. King 2005:54ff.).

Auch im Bereich der Gebrauchsqualität können bereits einige Testmöglichkeiten zugeordnet werden:

- Effizienz → *Benchmarktest*
- Produktivität → Testvalidierung
- Sicherheit
- Zufriedenstellung → Testvalidierung

Produktivität und Zufriedenstellung können wie bereits festgestellt nur im Rahmen einer tatsächlichen Anwendung getestet werden. Um die Gebrauchsqualität vorherzusagen, wird üblicherweise eine Evaluation der externen Qualität durchgeführt. Jedoch erfordern die Umstände einer Evaluation der externen Qualität in zahlreichen Fällen bereits eine Anpassung der Programme bzw. eine derart aufwendige Vorbereitung der Evaluation, dass fast eine reale Anwendungssituation nachgestellt wird, wie es die Evaluation der Gebrauchsqualität erfordert – diese beiden Evaluationsformen sind dementsprechend in einigen Fällen schwer abgrenzbar (vgl. King 2005:59).

Einige der Kriterien können – wie auch bezüglich der Merkmale festgestellt – negativ korrelieren: besonders anschaulich ist dies im Fall der drei Faktoren Qualität, Kosten und Zeit. Eine Maximalisierung der Qualität kann erhöhte Kosten und eine längere Laufzeit zur Folge haben, eine maximale Verkürzung der Laufzeit kann zu erhöhten Kosten und Abstrichen bei der Qualität führen – minimale Kosten können eine geringere Qualität und höhere Laufzeiten zur Folge haben (vgl. Hoffmann 2008:11f.). In solchen Fällen kann im Rahmen von spezifischen Evaluationen eine Priorisierung anhand der Anforderungen erfolgen. In allgemeinen Evaluationen kann die Priorisierung den Anwendern der Evaluation zur eigenen Gewichtung überlassen werden, hierfür sollten möglichst ausführliche, neutrale Informationen zur Entscheidungsfindung bereitgestellt werden.

4.4.5 Instrumente

Neben den Testmethoden spielen die verwendeten Instrumente und Materialien eine wichtige Rolle. Instrumente dienen der Datenerhebung und können mit beliebigen Testmethoden und weiteren Instrumenten kombiniert werden. Die Wahl der entsprechenden Instrumente hängt ebenfalls von den vorhandenen finanziellen und zeitlichen Möglichkeiten sowie von der Erfahrung der Evaluatoren und der eventuellen Beteiligung von Anwendern ab (vgl. EAGLES 1995:85). Testinstrumente, die während der Testphase verwendet werden, können von den Berichtsinstrumenten für die Auswertung unterschieden werden.

4.4.5.1 Testinstrumente

Ziel der Verwendung von Testinstrumenten ist es, aussagekräftige Daten zu sammeln. Dies kann auf manuelle oder automatische Weise geschehen: Zu den manuellen Instrumenten gehören Fragebögen (z.B. Umfragen, interaktive ins aktive Testverfahren integrierte Fragebögen, Debriefing-Fragebögen), Checklisten, Interviews, Beobachtungen (z.B. direkte Beobachtungen, Videoaufzeichnungen oder Beobachtungen durch Spiegelscheiben) und Think-Aloud-Protokolle. Die automatischen Instrumente sind Logging und Playback (vgl. EAGLES 1995:86-92). Beobachtungen und Think-Aloud-Protokolle sowie Logging und Playback werden in Evaluationen mit der Beteiligung von Anwendern verwendet. Die automatischen Instrumente sind besonders geeignet, um Daten zur Verständlichkeit und Erlernbarkeit zu erfassen (vgl. EAGLES 1995:91f.).

4.4.5.2 Berichtsinstrumente

In den verschiedenen Evaluationsphasen werden unterschiedliche Berichtsinstrumente verwendet, um Informationen für die Auswertung und Beurteilung zu sammeln, um die Evaluation zu dokumentieren, zu analysieren und an die Auftraggeber der Evaluation weiter zu geben. Es kann unterschieden werden zwischen Evaluationsbeschreibungen, Problemberichten, Ergebnisberichten und Beurteilungsberichten. Evaluationsbeschreibungen können sehr umfangreich sein und von den Gründen zur Durchführung der Evaluation über die Umgebungsfaktoren und Vorgaben bis zum Evaluationsablauf zahlreiche Informationen enthalten (vgl. EAGLES 1995:92-95). Im Problembericht werden die im Verlauf der Evaluation aufgetretenen Fehler und Probleme dokumentiert; diese Art von Bericht ist vor allem für entwicklungsorientierte Evaluationen von Bedeutung (vgl. EAGLES 1995:95). Ergebnisberichte enthalten die Ergebnisse der einzelnen Tests und Methoden. Sie werden in jenen Evaluationen benötigt, in denen keine Checklisten oder ähnlichen Testinstrumente verwendet werden, um eine Dokumentation der Ergebnisse sicherzustellen (vgl. EAGLES 1995:95f.). Beurteilungsberichte enthalten Zusammenfassungen des Ergebnisses (und eventuell Empfehlungen) im Bezug zu den in der Anfangsphase definierten Zielen und Anforderungen. Sie stellen das Hauptdokument zum Bericht der Evaluation an den Auftraggeber dar und werden von anderen Berichtsinstrumenten im Anhang ergänzt (vgl. EAGLES 1995:96f.).

4.4.6 Testmaterial

Wie bereits in Kap. 4.4.3.1 beschrieben, sind die Repräsentativität und der Realitätsgrad wichtige Eigenschaften des Testmaterials. Zum Testen können drei Arten von Material verwendet werden: Testsätze (*test sets*), unbearbeitetes Material natürlichen Textes; Testreihen

(*test suites*), bearbeitetes Material zum Testen bestimmter Phänomene; und Testkollektionen (*test collections*), Testmaterial mit bereits vorbereiteten Ergebnissen.

Im Fall von unbearbeitetem Material ist der Realitätsgrad sehr hoch, es muss jedoch besonderes Augenmerk auf die Repräsentativität gelegt werden. Jeder Text kann repräsentativ für ein bestimmtes Phänomen oder einen Bereich sein, es darf nur nicht davon ausgegangen werden, dass jeder beliebige Text allgemein repräsentativ ist (vgl. EAGLES 1995:36f.).

Bearbeitetes Material ist ebenfalls mit Vorsicht zu behandeln, da die Repräsentativität der Phänomene subjektiv beeinflusst sein kann und der Realitätsgrad stark beeinträchtigt ist. Die Bearbeitung von Material ist darüber hinaus aufwendig, sowohl in der Erstellung als auch in der Auswertung (vgl. Sparck-Jones & Galliers 1995:139). Es muss darauf geachtet werden, das Material nicht auf ein bestimmtes Programm auszurichten (vgl. Sparck-Jones & Galliers 1995:140).

Vorbereitete Lösungen in Form von Testkollektionen sind nicht nur sehr kosten- und zeitaufwendig in der Erstellung, sie sind außerdem für einige Programmtypen ungeeignet: bei der Evaluation von Programmen der maschinellen Übersetzung und auch bei der Evaluation von *translation memories* hat sich gezeigt, dass es nicht möglich war, erwartete Lösungen festzulegen (vgl. Sparck-Jones & Galliers 1995:148 & EAGLES 1995:153). Ähnliches gilt für TEP: die unterschiedlichen Ansichten und Erwartungen bei der Extraktion von Termini wurden bereits in Kapitel 1.1 beschrieben und würden kaum eine allgemeingültige Festlegung erwarteter Lösungen ermöglichen. Die Verwendung von multilingualem Testmaterial ist besonders problematisch: in verschiedenen Sprachen ist die Vergleichbarkeit schwer zu gewährleisten (vgl. Sparck-Jones & Galliers 1995:144).

Nachdem nun die einzelnen Bestandteile der Methodologie vorgestellt wurden, soll der Evaluationsablauf als vollständiger Prozess zusammengefasst werden.

4.5 Ablauf einer Evaluation

Bei der Entwicklung eines Evaluationsdesigns kann eine systematische und schrittweise Vorgangsweise unterstützend wirken (vgl. Sparck-Jones & Galliers 1995:194). Aus diesem Grunde soll hier nach der grundlegenden Betrachtung des EAGLES-Evaluationsrahmens der in Kapitel 4.1.1.2 beschriebene Evaluationsprozess gemäß ISO-IEC 14598 nochmals kurz mit den laut EAGLES-Evaluationsrahmen vorgesehen Schritten zusammengefasst werden:

- *Definition der Evaluationsanforderungen*:
  Analyse der Evaluationssituation, des Evaluationsziels, der potentiellen Anwender und Käufer, der Setups und Umgebungsfaktoren;
  Erstellung eines Prozessmodells, Erweiterung des Prozessmodells und Integration der funktionalen und nicht funktionalen Anforderungen.

- *Spezifizierung der Evaluation*:
  Umwandlung in wesentliche informative Merkmale, Definition von zugehörigen Werten und Maßen, Suche nach passenden Methoden.
- *Design der Evaluation*:
  Entscheidung für investigativen oder experimentellen Ansatz, Wahl eines oder mehrerer Teststypen, Festlegung der Methoden, einschließlich der zu verwendenden Materialien und Interpretationsvorgaben, Bestimmung der Instrumente.
  (Vorbereitung: ggf. Suche der Anwender bzw. weiteren Personals und Schulung, Erstellung des Materials, Vorbereitung der Instrumente, Probeläufe.)
- *Durchführen der Evaluation*:
  Messung, Gewichtung der Ergebnisse nach Vorgaben und Grenzwerten, Bewertung nach Interpretationsvorgabe, Redaktion der Berichte.

Wichtig für die Durchführung der Evaluation ist eine gründliche Vorbereitung. Zur Vorbereitungsphase gehören neben der Spezifikation der Anforderungen, Merkmale, Maße und Methoden (einschließlich des Testmaterials und den Vorgaben zur Interpretation), Testabläufen (einschließlich der zu verwendenden Instrumente) eventuell auch Probeläufe und Probetests. Neue bzw. unbekannte Methoden sollten im Vorfeld ausprobiert und automatisierte Verfahren auf ihre Funktionstüchtigkeit getestet werden. Es sollte außerdem darauf geachtet werden, dass das im schlussendlichen Testlauf verwendete Testmaterial in der Vorbereitungsphase noch nicht verwendet wurde (vgl. Sparck-Jones & Galliers 1995:154).

In der Durchführungsphase sollten die einzelnen Schritte nicht miteinander vermischt werden. Es ist wichtig die Reihenfolge einzuhalten: zuerst messen, anschließend gemäß der Vorgaben und Grenzwerte gewichten und erst zum Schluss gemäß der Interpretationsvorgabe anhand der zu Beginn definierten Kriterien bewerten (vgl. EAGLES 1995:148). Für den Fall, dass im Evaluationsverlauf Funktionalitäten entdeckt werden, die dem Evaluator in der Planungsphase nicht bekannt waren, kann das Schema unter Umständen auch während der Evaluation angepasst werden (vgl. EAGLES 1999).

4.6 Zusammenfassung

Im Rahmen dieses Kapitels wurde die Evaluation von sprachverarbeitenden Programmen mit einem besonderen Schwerpunkt auf dem EAGLES-Evaluationsrahmen betrachtet. Nach einer Einführung in die Konzepte und Grundlagen der Softwareevaluation und Softwarequalität wurden die Grundzüge des EAGLES-Evaluationsrahmens präsentiert. Anschließend wurde der Prozess zur Entwicklung einer Evaluationsmethodologie anhand der Ausgangssituation und den jeweiligen Anforderungen über die Definition von Merkmalen und Maßen bis zum Evaluationsdesign einschließlich der Entwicklung von Methoden, Festlegung von Testtypen,

Instrumenten und Testmaterialien beschrieben. Eine Übersicht über den Ablauf einer Evaluation bildet den Abschluss dieses Kapitels.

Im nächsten Kapitel werden bereits erfolgte Evaluationen von Terminologieextraktionsprogrammen und ihre Methodologie betrachtet und Ideen zu einer Weiterentwicklung der Evaluationsmethodologie vorgestellt.

5 Evaluationen von Terminologieextraktionsprogrammen

Nach einer umfassenden theoretischen Betrachtung der Entwicklung von Evaluationsmethodologien auf der Grundlage des EAGLES-Evaluationsrahmens sollen nun Evaluationen von Terminologieextraktionsprogrammen analysiert werden. In Kapitel 5.1 wird ein Überblick über bisherige Evaluationen von TEP und ihre Methodologien und Testmaterialien gegeben. In Folge soll der methodologische Schwerpunkt zur Qualität der Identifikationsergebnisse diskutiert werden (Kap. 5.2) und näher auf zwei umfassendere Evaluationsmethodologien eingegangen werden (Kap. 5.3). Abschließend werden einige Ideen zur Weiterentwicklung der Evaluationsmethodologie von TEP präsentiert (Kap. 5.4) sowie konkrete Vorgaben und Methoden entwickelt (Kap. 5.5).

5.1 Forschungsüberblick

Wie bei allen Arten von Software kann auch für Terminologieextraktionsprogramme keine allgemeine Aussage getroffen werden, welches TEP das „beste“ ist (vgl. Zerfaß 2006:24). In Kapitel 2 wurden die Unterschiede von TEP betrachtet, die sich nicht nur auf offensichtliche Aspekte wie z.B. die zu verarbeitenden Sprache beziehen, sondern auch auf das Design und die Implementierung der Methoden, den Identifikationsansatz, die Eigenständigkeit bzw. die Integration in ein umfassenderes Programm (vgl. L’Homme *et al.* 1996:292). Diese Vielfalt spiegelt sich auch in der Evaluationsmethodologie wider: da TEP oft für spezifische Anwender entwickelt und auch evaluiert werden, herrscht wenig Konsens im Bereich einer möglichen Standardisierung der Evaluationsmethodologie (vgl. Vivaldi & Rodríguez 2007:225). Im Jahr 2001 stellten Ahmad & Rogers fest, dass die Evaluierung von TEP noch in den Kinderschuhen steckt (vgl. 2001:751), und auch 2002 beklagte Lieske einen Mangel an komparativen Evaluationen für TEP. Er begründete dies mit der großen methodologischen Herausforderung eine komparative Evaluation von TEP zu planen und durchzuführen (vgl. Lieske 2002:109).

In diesem Kapitel sollen einige Forschungsarbeiten zur Evaluation von TEP vorgestellt werden. Es ist selbstverständlich unmöglich sämtliche Arbeiten des Gebietes abzudecken. Da das Ziel der Arbeit die Entwicklung einer anwenderorientierten Evaluationsmethodologie ist, werden besonders viele Arbeiten mit terminologischem Schwerpunkt und weniger Arbeiten mit computerlinguistischem Schwerpunkt berücksichtigt. Letztere konzentrieren sich zumeist auf entwicklerorientierte Glass-Box-Evaluationen, die zwar interessante Informationen enthalten können, aber nicht dem priorisierten Methodologiebereich angehören. Der Schwerpunkt dieses Kapitels liegt darüber hinaus auf der Methodologie und nicht auf den konkreten

Evaluationsergebnissen[28]. Es werden nun kurz die einzelnen Publikationen nach terminologischem, translatorischem bzw. computerlinguistischem Schwerpunkt geordnet vorgestellt.

5.1.1 Aus terminologischer Sicht

Zur terminologischen Sicht werden all jene Arbeiten gezählt, die einen eindeutig terminologischen Schwerpunkt aufweisen, auch wenn die Verfasser eventuell Translatoren sind oder anderen Forschungsbereichen angehören. Ein gemeinsames Merkmal der terminologischen Arbeiten ist der hohe Anteil an Black-Box-Evaluationen: bis auf einige in der Publikation von Cabré *et al.* 2001 angeführten Untersuchungen, basieren sämtliche Arbeiten dieses Abschnitts auf Black-Box-Verfahren. Die Arbeiten werden in chronologischer Reihenfolge präsentiert.

Bereits 1996 entwickelten L'Homme *et al*. ein vielzitiertes Evaluationsraster für Terminologieextraktionsprogramme. Es besteht aus zwei Phasen der Evaluation: Vor- und Hauptevaluation. Im Rahmen der Vorevaluation wird die generelle Eignung des Programms anhand verschiedener Kriterien (Art des Programms, angebotene Sprachen, Identifikationsmethode, Art der extrahierten potentiellen Termini sowie verschiedene Parameter, Exportmöglichkeiten und Arten der Ergebnisrepräsentation) überprüft (vgl. L'Homme *et al.* 1996:294f.). Die Hauptevaluation findet nach dem Extraktionsvorgang statt und bezieht sich auf die Qualitätsanalyse des Identifikationsergebnisses anhand von Precision- und Recall-Messungen (vgl. L'Homme *et al.* 1996:302).

Cabré *et al.* publizierten 2001 einen Überblick über verschiedene TEP, in dem sie die jeweiligen Programme und ihre Identifikationsmethoden kurz vorstellten (vgl. Cabré *et al.* 2001:56). Im Anschluss an die Präsentation jedes Programms erfolgte eine kurze Evaluierung der Stärken und Schwächen desselben auf der Grundlage von zugehörigen Publikationen und am Ende des Artikels werden in einer strukturierten Übersicht einige Charakteristika der einzelnen Programme (Verwendung linguistischer Hilfsmittel, Strategien der Segmentierung, Strategien der Filterung, Klassifizierung der potentiellen Termini und Qualität des Identifikationsergebnisses) verglichen (vgl. Cabré *et al.* 2001:79ff.). Von den Autoren wurden die Programme nicht evaluiert. Die „Evaluation" beschränkt sich auf die Wiederholung von Zahlen zur Qualität des Identifikationsergebnisses und Ideen zu möglichen Ursachen für gute bzw. schlechte Ergebnisse sowie eine Beurteilung der Möglichkeiten zur Validierung (vgl. Cabré *et al.* 2001:57-79). Cabré *et al.* schlussfolgern, dass das Ziel darin bestehe, die Precision-Werte zu erhöhen, und die Silence-Werte zu reduzieren (vgl. Cabré *et al*. 2001:84).

Lieske beschreibt 2002 eine pragmatisch ausgerichtete Methodologie zur vergleichenden Evaluation von Terminologieextraktionsprogrammen, die für ein Unternehmen ent-

[28] Die jeweiligen Evaluationsergebnisse sind in den entsprechenden Arbeiten enthalten.

wickelt und angewendet wurde (vgl. Lieske 2002:109). Es handelt sich um einen ganzheitlichen Evaluationsansatz, der auf dem EAGLES-Evaluationsrahmen und der ISO-IEC Norm 9126 aufbaut (vgl. Lieske 2002:115) und anhand dessen fünf TEP mithilfe von Texten in vier Sprachen und verschiedenen Beteiligten (u.a. neun Validierer) evaluiert wurden (vgl. Lieske 2002:123). In der Publikation wurden Grundlagen für die Anforderungen, Auszüge aus Merkmal- und Maßlisten, verwendeten Tests und Daten, Informationen zum Ablauf und zu den Interpretationsvorgaben dokumentiert. Für die Evaluation wurde eine Validierungsumgebung für Termini entwickelt (vgl. Lieske 2002:114-121). Auf die Methodologie von Lieske wird in Kapitel 5.3 näher eingegangen.

Gasser suchte im Jahr 2004 nach einem effizienten Verfahren zum Befüllen einer Terminologiedatenbank und evaluierte in diesem Rahmen die mono- und bilinguale Extraktion mit einem TEP im Vergleich zur manuellen Extraktion. Außerdem sollten im Rahmen der Evaluation geeignete Parametereinstellungen für die TEP gefunden werden. Gassers Evaluation sollte als erste Stufe einer Kosten-Nutzen-Analyse eines Unternehmens dienen (vgl. Gasser 2004:5), die allem Anschein nach mit der Arbeit von Fleischmann 2006 fortgeführt wurde. Schwerpunkte der Evaluation liegen im Versuch durch eine geeignete Parameterwahl den Silence-Wert zu reduzieren und durch bilinguale Evaluation auch für Assistenten eine Zuordnung von Synonymen zu ermöglichen (vgl. Gasser 2004:6).

Fleischmann untersuchte 2006, ob computergestützte TE wirtschaftlicher ist als manuelle TE und orientierte sich hierbei am Verhältnis zwischen Qualität und Wirtschaftlichkeit (vgl. Fleischmann 2006:5). Im Zuge der Evaluierung wurden vier TEP verglichen, ein besonderer Schwerpunkt lag auf der Effizienz. Nur ein Teil des Korpus wurde manuell extrahiert, das Ergebnis als Benchmark für die TEP verwendet. Außerdem wurde ein Vergleich der vollständigen validierten Ergebnisse der TEP durchgeführt (vgl. Fleischmann 2006:11f.). Die Qualität des Identifikationsergebnisses dominierte neben der Nachbearbeitung, Validierung und Anpassbarkeit diese Evaluation (vgl. Fleischmann 2006:109-119), obwohl im Vorfeld weit umfassendere Kriterien definiert wurden (vgl. Fleischmann 2006:39-41, 45).

Im Jahr 2007 evaluierte Wanhoff ein für die deutsche Sprache entwickeltes TEP im Bezug auf seine Grenzen und Möglichkeiten bei der Extraktion aus dem Italienischen, und versuchte die Anwendung zu optimieren. Zur weiteren Optimierung gab sie einen Überblick über Identifikationsmethoden anderer TEP und analysierte die Wortbildung im Italienischen (vgl. Wanhoff 2007:6). Die Evaluation beschränkt sich auf die Analyse der Silence- und Noise-Werte unter besonderer Berücksichtigung von Wortbildungsmustern (vgl. Wanhoff 2007:33).

5.1.2 Aus translatorischer Sicht

Translatoren (überwiegend Übersetzer), die sich mit TEP auseinandersetzen, tun dies oft aus terminologischer Sicht – einige der im obigen Kapitel vorgestellten Autoren haben ein Übersetzerstudium absolviert. Dies mag einer der Gründe sein, aus dem nur wenige Publikationen mit rein translatorischem Schwerpunkt zu diesem Thema existieren. Übersetzer sehen Terminologieextraktionsprogramme als Hilfestellung im Prozess der Vorbereitung einer Übersetzung, und nicht als Programme zur Unterstützung ihres gesamten Arbeitsablaufes. Im Moment, in dem die Terminologiearbeit ihren Alltag dominiert, werden sie zu Terminologen. Dennoch gibt es zwei Beispiele für Evaluationen aus Übersetzersicht: Beim ersten handelt es sich um die vage Beschreibung der Evaluation eines TEP, die von mehreren Übersetzerinnen durchgeführt wurde, und weiteren Übersetzern in einer Fachzeitschrift empfohlen wird – das zweite Beispiel ist die Arbeit von Saß (2004).

Zerfaß berichtet von einer Übersetzerin, die eine vergleichende Evaluation zweier TEP durchgeführt hat. Zunächst wurde eine manuelle Referenzliste der vorhandenen potentiellen Termini erstellt, die Extraktion wurde mit beiden Programmen durchgeführt und die Ergebnisse in Excel-Listen mithilfe von Formeln verglichen. Dieser Abgleich konnte keine potentiellen Termini berücksichtigen, die durch fehlende Grundformenbildung bzw. Rechtschreibfehler geringfügig unterschiedlich waren, auch getrennte Wörter führten nicht zu Übereinstimmungen. Zerfaß hat diese Vorgangweise selbst ebenfalls angewendet und berichtet über lange Verarbeitungszeiten und aufwendige Datenkonvertierung sowie mangelnde Vergleichbarkeit (vgl. Zerfaß 2006:24). Hier soll nur angemerkt werden, dass eine einzelne Evaluation natürlich alles andere als repräsentativ ist. Leider konnten nur eine weitere Evaluation mit translatorischem Schwerpunkt ausfindig gemacht werden.

Im Jahr 2004 führte Saß eine umfassende Evaluation von vier TEP zur monolingualen und drei TEP zur bilingualen Terminologieextraktion durch. Die Evaluationsmethodologie wurde anhand des EAGLES-Evaluationsrahmens und der ISO-IEC Norm 9126 entwickelt (vgl. Saß 2004:17). Als Daten dienten zwei Beispieltexte und zum Vergleich manuell extrahierte Ergebnisse (vgl. Saß 2004:27). Die Evaluationskriterien umfassten zahlreiche in Oberkategorien eingeteilte Aspekte (Technische Beschreibung, Bedienungsoberfläche, Terminologieextraktion, Eingabe und Import, Ausgabe und Export, Validierung der Ergebnisse und ökonomische Aspekte (vgl. Saß 2004:29), u.a. fand eine detaillierte Auswertung der Qualität des Identifikationsergebnisses anhand von Recall- und Precision-Werten unter Beachtung der Auswirkung unterschiedlicher Parametereinstellungen statt (vgl. Saß 2004:75-88). Auf die bilinguale Extraktion wird gesondert eingegangen (vgl. Saß 2004:22, 25 & 87f.). Saß Methodologie wird ebenfalls in Kapitel 5.3 behandelt.

5.1.3 Aus computerlinguistischer Sicht

Eine sehr technische Perspektive der Evaluation von TEP ermöglicht die Arbeit von Kageura *et al.* 2001. Es handelt sich um die Dokumentation eines TEP-bezogenen Teils des NTCIR-Workshops[29]. Eine der Teilaufgaben behandelte die Terminologieextraktion. In der Publikation werden die Wettbewerbsbedingungen beschrieben (vgl. Kageura *et al.* 2001:155-162) und kurz die Aufsätze der einzelnen teilnehmenden Teams zusammengefasst (vgl. Kageura *et al.* 2001:163-168). Das Testmaterial wurde in roher und annotierter Form bereitgestellt, um keine Methode der Vorverarbeitung oder Identifikation zu benachteiligen. Zu Diskussionszwecken wurden vorher bestimmte potentielle Termini festgelegt, die jedoch weder eine Benchmark darstellen, noch als repräsentativ angesehen werden (vgl. Kageura *et al.* 2001: 157ff.). Kageura *et al.* haben keine eigene Evaluation durchgeführt.

Im Jahr 2005 publizierte Witschel eine umfassende Dokumentation der Entwicklung eines eigenen TEP, die auch die Evaluation des selbst entwickelten Programmes beinhaltete. Es handelt sich um eine Glass-Box-Evaluation des Gesamtprogramms (vgl. Witschel 2005: 73ff.) sowie seiner einzelnen Komponenten (Differenzanalyse, verschiedene Relevance Feedback-Methoden, vgl. Witschel 2005:75-95). Als Maße wurden Recall, Precision und der F-Wert verwendet (vgl. Witschel 2005:71). Das Testmaterial bestand aus zwei Texten aus unterschiedlichen Fachgebieten und zugehörigen vorbereiteten Ergebnissen, die manuell unter Berufung auf Lexika zur Abgrenzung Fachsprache/Gemeinsprache erstellt wurden (vgl. Witschel 2005:72f.). Darüber hinaus wurde ein Überblick über die Ergebnisse anderer Programme gegeben, diese wurden jedoch nicht von Witschel getestet – die Ergebnisse entstammten anderen Publikationen (vgl. Witschel 2005:95f.).

Vivaldi & Rodríguez publizierten 2006 & 2007 Aufsätze über eine Evaluation ihres selbst entwickelten TEP. Die von ihnen entwickelte Glass-Box-Methodologie (vgl. Vivaldi & Rodríguez 2007:237) soll auch für andere TEP genutzt werden können: es werden sowohl die Termini (Qualität des Identifikationsergebnisses) als auch das TEP evaluiert (vgl. Vivaldi & Rodríguez 2006:12). Als Maße für die Qualität des Identifikationsergebnisses dienten Precision, Recall sowie akkumulierte Kombinationen dieser und anderer Werte. Vivaldi & Rodríguez sind sich allerdings der Einschränkungen bei der Definition „korrekter“ Termini und somit des Recalls sowie der Tatsache, dass für die Verwendung dieser Maße ein gewisses Schwarz-Weiß-Denken angenommen werden muss, bewusst (vgl. 2006:12). Besonders betont werden der Wunsch nach größerer Standardisierung sowie die Bedeutung der Evaluation innerer Komponenten (vgl. Vivaldi & Rodríguez 2006:13). Zusätzlich zum computer-

[29] NTCIR steht für *Nacsis Test Collection for Information Retrieval*. Der NTCIR-Workshop wird seit 1999 vom *National Institute of Informatics* (früher *National Center for Science Information Systems*) in Tokio organisiert (vgl. Kageura et al. 2001:155).

linguistischen Hintergrund und Ansatz vertreten Vivaldi & Rodríguez eine stark terminologische Sichtweise und evaluieren im Hinblick auf systematische Terminologiearbeit (vgl. 2007: 226). Ihre Evaluationsmethodologie ist sehr an der Funktionalität ausgerichtet und soll kontextunabhängige Evaluationen von kontextunabhängigen TEP ermöglichen (vgl. Vivaldi & Rodríguez 2007:237) – dies entspricht einem der von King identifizierten Paradigma, in dem lediglich die kontextunabhängige Funktionalität evaluiert wird (vgl. King 2005:46ff.).

Diese drei Arbeiten stellen nur einen sehr kleinen Ausschnitt aus vorhanden Publikationen zur Evaluation von Terminologieextraktionsprogrammen aus computerlinguistischer Sicht dar – für die Zwecke dieser Arbeit sollte die Auswahl jedoch ausreichend sein. Interessant ist in diesem Zusammenhang die Empfehlung des EAGLES-Projektes, dass Entwickler nicht ihre eigene Software testen sollten (vgl. EAGLES 1995:67) – in der Praxis ist allem Anschein nach das Gegenteil der Fall.

5.1.4 Methodologie

In diesem Unterkapitel soll die Methodologie der bisherigen Evaluationen von TEP im Hinblick auf einige bedeutende Aspekte wie Zielsetzung und Planung, Evaluationskriterien und Methoden untersucht werden. Das verwendete Testmaterial sowie die als Ergebnis der Evaluationen gezogenen Schlussfolgerungen zur Vergleichbarkeit und Standardisierung werden in den darauffolgenden Unterkapiteln behandelt. Auf die beiden Arbeiten, die auf dem EAGLES-Evaluationsrahmen aufbauen (Lieske 2002 & Saß 2004), soll hier nicht näher eingegangen werden, sie werden in Kapitel 5.3 im Detail vorgestellt.

5.1.4.1 Zielsetzung & Planung

Unbestritten ist die Planung ein wichtiger Aspekt einer Evaluation. Bereits im Vorhinein muss das Ziel der Evaluation definiert werden. Dies bezieht sich nicht nur auf das Evaluationsziel, es muss darüber hinaus ein Extraktionsziel festgelegt werden: die Frage, wonach gesucht wird bzw. welche Termini gefunden werden sollen, muss eindeutig beantwortet werden (vgl. Witschel 2005:16). Von den Zielen hängen zahlreiche spätere Faktoren ab, z.B. die weitere Entwicklung der Methodologie vom Evaluationsziel (vgl. Kageura *et al.* 2001:154f.), bzw. die Bewertung der Brauchbarkeit eines Programmes vom Extraktionsziel (vgl. Zerfaß 2006:25).

Die Anwenderabhängigkeit wird in mehreren Arbeiten thematisiert: Heid verweist auf diverse Anwendergruppen, die mit ihren unterschiedlichen Bedürfnissen allgemeine Evaluationen von TEP erschweren, und schlägt als Lösung vor, weniger umfassende Evaluationen anzustreben (vgl. 2006:98). Witschel beschreibt das anwenderabhängige Verständnis von Relevanz (vgl. 2005:16), und auch Kageura *et al.* weisen darauf hin, dass unterschiedliche

Ziele und Kriterien bei der Interpretation der Ergebnisse beachtet werden sollen und keine allgemeinen Urteile über die Eignung von Programmen getroffen werden können (vgl. 2001: 154f.).

Diversität wird auch für die Evaluationskriterien gefordert: laut L'Homme *et al.* sollten Anwender mehrere Faktoren beachten, bevor sie ein Programm kaufen (vgl. 1996:292). Mit unterschiedlichen Kriterien befasst sich das folgende Unterkapitel.

5.1.4.2 Evaluationskriterien

Die Erfassung der Kriterien zur Evaluation von TEP ist unterschiedlich. Falls ein bestimmtes Unternehmen als Auftraggeber fungiert, werden in manchen Fällen direkt vom Unternehmen formulierte Anforderungen beschrieben und gegebenenfalls erweitert (vgl. Fleischmann 2006:69). L'Homme *et al.* geben an, für die Entwicklung des Evaluationsrasters auf Literatur und die Erfahrung im Umgang mit zwei TEP zurückgegriffen zu haben (vgl. 1996:292). Zielinski & Ramírez-Safar führten für ihre 2005 und 2006 publizierten Arbeiten keine Evaluation, sondern eine Umfrage durch, in der unter anderem Anforderungen an Terminologieextraktionsprogramme erhoben wurden. Auf die Ergebnisse soll in Kapitel 5.4.2 und 5.4.3 näher eingegangen werden. In diesem Kapitel wird ein Überblick über jene Kriterien gegeben, die in den bisherigen Evaluationen vorgeschlagen wurden, bzw. angewendet worden sind.

In fast jeder Evaluation von Terminologieextraktionsprogrammen wird in irgendeiner Form die Qualität des Identifikationsergebnisses behandelt, zumeist anhand der in Kapitel 2.4.1 vorgestellten Wertepaare Precision und Recall bzw. Noise und Silence (vgl. Zielinski & Ramírez-Safar 2005:3). Dieser Schwerpunkt im Rahmen der Evaluationen wird in Kapitel 5.2 näher behandelt.

Es gibt allerdings auch in bisherigen Evaluationen weitere Kriterien, die zur Beurteilung verwendet werden. Viele dieser Kriterien beziehen sich ebenfalls auf die Qualität des Identifikationsergebnisses oder auf die verwendeten Identifikationsmethoden, die wiederum Einfluss auf die Qualität des Identifikationsergebnisses haben, z.B. Mindestfrequenzwerte, Umgang mit Sonderzeichen, Umgang mit unbekannten Token, etc. (vgl. L'Homme *et al.* 1996:300f.). Zu diesen Kriterien gehören auch einige der Kategorien im Überblick von Cabré *et al.*, z.B. Strategien der Tokenisierung, Filterung (vgl. 2001:79f.).

Darüber hinaus gibt es Kriterien, die keinen direkten oder indirekten Bezug zur Qualität des Identifikationsergebnisses haben, wie z.B. die verfügbaren Sprachversionen eines TEP, bzw. die Erfordernis, dem Programm weitere linguistische Daten zur Verfügung zu stellen (vgl. Cabré *et al.* 2001:79). Ebenso fallen in diesen Bereich die mögliche Kompatibilität des TEP mit anderen Programmen, die verwendbaren Formate, Import- und Exportmöglichkei-

ten, Interaktionsmöglichkeiten in verteilten Workflows, Bedienerfreundlichkeit sowie Möglichkeiten zur Nachbearbeitung (vgl. Zielinski & Ramírez-Safar 2006:28, Fleischmann 2006: 40f.).

Weitere Kriterien sind Zeit- und Kostenfaktoren: diese basieren auf Werten wie der Laufzeit und dem Arbeitsaufwand sowie Kosten für den Programmerwerb und die Bezahlung von Arbeitskräften. Einige dieser Kriterien sind wiederum eng verbunden mit der bereits genannten Qualität des Identifikationsergebnisses (vgl. Fleischmann 2006:39f.).

Die Anpassung der Anforderungen ist in zahlreichen Publikationen erfolgt, L'Homme *et al.* verweisen z.B. explizit auf die unterschiedliche Anforderungen von Übersetzern und Terminologen (vgl. 1996:295f.) und darauf, dass Evaluationskriterien laufend an neue Entwicklungen angepasst werden müssen (vgl. 1996:309).

5.1.4.3 Methoden

Entsprechend der auf Qualität des Identifikationsergebnisses fokussierten Kriterien dient auch die Mehrzahl der angewandten Methoden Precision- und Recall- bzw. Silence- und Noise-Messungen bzw. Berechnungen. Zwei klassische Methoden treten besonders häufig auf: ein Benchmarktest anhand einer Testkollektion zur Erfassung des Recall- bzw. Silence-Wertes sowie eine vergleichende Analyse zur Berechnung des Precision- oder Noise-Wertes.

Für die Testkollektion des Benchmarktests (der auch vergleichend durchgeführt werden kann) werden potentielle Termini aus einem Teil des Korpus (nur in seltenen Fällen aus dem gesamten Korpus, vgl. Witschel 2005:72f., Wanhoff 2007:26f.) zumeist manuell extrahiert, und das Ergebnis als Benchmark für die Berechnung des Recalls des Ergebnisses eines oder mehrerer TEP verwendet (vgl. Fleischmann 2006:58f. & 112-115, L'Homme 1996:302f.). Alternativ zur manuellen Extraktion können die Testkollektionen auch computergestützt erstellt werden: Witschel sieht für eine Baseline all jene Wörter und Phrasen als potentielle Termini an, die bei einer maschinellen Übersetzung nicht bzw. falsch übersetzt worden sind, da sie nicht im systemeigenen Wörterbuch enthalten sind (vgl. Witschel 2005:72).

Wanhoff untersucht im Anschluss an die Berechnungen die fehlenden Termini auf besondere Charakteristika bzgl. der Wortbildungsmuster, um Gründe für die fehlende Identifikation ausfindig zu machen (vgl. Wanhoff 2007:25).

Für die vergleichende Analyse werden mehrere TEP auf den gesamten Korpus angewendet, die Ergebnisse validiert und anschließend Berechnungen der Precision- bzw. Noise-Werte durchgeführt und verglichen (vgl. Fleischmann 2006:11 & 110-112). In einigen Fällen werden bei vergleichenden Analysen anhand der unterschiedlichen Ergebnisse Benchmarks errechnet bzw. festgelegt und im Vergleich zu diesen Recall-Berechnungen durchgeführt. Zum Zwecke einer detaillierteren Perspektive auf die Qualität des Identifikationsergebnisses

verwendet Fleischmann bei der Validierung fünf Kategorien der Korrektheit im Gegensatz zur üblichen Einteilung in korrekt/inkorrekt (vgl. 2006:52ff.).

Die Einstellmöglichkeiten der Parameter zur Verbesserung der Qualität des Identifikationsergebnisses werden von einigen Evaluatoren durch Ausprobieren verschiedener Einstellungen untersucht. Dies geschieht z.B. durch Probevalidierungen der Ergebnisse vor der eigentlichen Extraktion zur Analyse der allgemeinen Qualität des Identifikationsergebnisses, um im Hauptextraktionsdurchgang bereits die optimalen Einstellungen zu verwenden (vgl. Gasser 2004:29-32); oder durch eine systematische schrittweise Veränderung der Parameter nach der bereits durchgeführten Extraktion und Analyse des Identifikationsergebnisses, um die Qualität noch zu verbessern, bzw. weil in der eigentlichen Extraktion Standardwerte verwendet werden sollen, um das Ergebnis von vergleichenden Analysen nicht zu beeinflussen (vgl. Witschel 2005:73).

Zur Kosten-Nutzen-Rechnung wird zumeist der Zeitaufwand zur Erstellung einer validierten Terminologieliste (bzw. -datenbank) mithilfe manueller Extraktion mit dem Zeitaufwand des Extraktionsvorgangs bei der computergestützten Extraktion verglichen. Die Argumentation und Berechnungsart ist unterschiedlich präzise und verständlich dokumentiert: Gasser rechnet anhand von Faustregeln und Annahmen und vergleicht die computergestützte mit der manuellen Extraktion, obwohl sich diese auf eine Stichprobe beschränkt und der Zeitaufwand nicht vollständig gemessen, sondern überschlagen wurde (vgl. 2004:36ff.); Fleischmanns Berechnungen sind detaillierter dokumentiert und enthalten Vergleichszahlen zur manuellen Extraktion nur dann, wenn sie auch tatsächlich erfasst wurden (vgl. 2006:116-122).

Vivaldi & Rodríguez beschreiben Initiativen zur Verwendung von möglichst neutralen und objektiven Methoden, z.B. im Rahmen von Evaluationscontests, durch die Bereitstellung von annotierten Korpora sowie durch die Standardisierung von Maßen, zur Berechnung der Übereinstimmung multipler Bewertungen und Methoden zur Validierung von Ergebnissen (vgl. 2007:228). Sie selbst verwenden einige dieser Methoden, u.a. die Berechnung der Übereinstimmung multipler Bewertungen im Rahmen von Identifikationsergebnisvalidierungen (vgl. 2007:238).

5.1.5 Testmaterial

Das Testmaterial hat starken Einfluss auf den Evaluationsverlauf, insbesondere auf die Qualität des Identifikationsergebnisses, die in zahlreichen Evaluierungen im Vordergrund steht (vgl. L'Homme 1996:303). Ahmad & Rogers fordern für jede Evaluation eines TEP die Einbeziehung des verwendeten Korpus in die Evaluation – eine Analyse seiner Struktur und Verarbeitung (vgl. 2001:727). In den meisten Arbeiten wird zumindest das als Korpus ver-

wendete Material beschrieben. In diesem Kapitel werden die Themen Korpusgröße, Testkollektionen und Vorverarbeitung besprochen.

5.1.5.1 Korpusgröße

Wie bereits in Kapitel 2.4.3 festgestellt, benötigen statistische Methoden zur Terminologieidentifikation ausreichend umfangreiche Korpora, um qualitativ hochwertige Ergebnisse liefern zu können (vgl. Fleischmann 2006:52). Die Evaluation des Identifikationsergebnisses im Hinblick auf Vollständigkeit, erfordert zur Überprüfung die Erstellung einer Testkollektion: einer Liste potentieller Termini, die im entsprechenden Text bzw. Korpus enthalten sind. Da die Erstellung einer solchen Liste aufwendig ist, werden die Korpora in einigen Fällen klein gehalten, um den Aufwand zur Erstellung einer Testkollektion zu reduzieren (vgl. Wanhoff 2007:26). Eine andere Möglichkeit besteht darin, nur Teile des Gesamtkorpus manuell zu extrahieren (z.B. Fleischmann 2006:58f., Gasser 2004:38). Kageura *et al.* weisen jedoch darauf hin, dass bei kleinen Korpora die rein numerische Betrachtung der Qualität des Identifikationsergebnisses (z.B. Recall) missinterpretiert werden könnte, da die Zahlen nicht unbedingt zuverlässig sein müssen (vgl. 2001:158). Eine solche Fehlinterpretation lässt z.B. das Ergebnis von Gassers Stichprobe vermuten: sie erhielt einen Recall-Wert von 100% nach der manuellen Extraktion einer „Stichprobe“ im Umfang von 10 Seiten und stellte dieses Ergebnis nicht in Frage, sondern verwendete es als Grundlage für weitere Berechnungen (vgl. Gasser 2004:38).

Außerdem kann es passieren, dass die statistischen Funktionen eines Programmes durch Stückelung des Korpus nicht vollständig genutzt werden: Gasser verwendete für die Extraktion ein TEP, das rein statistische Identifikationsmethoden verwendet (vgl. Gasser 2004:13). Um die Effektivität der Ausschlussdatenbank zur Noise-Reduktion zu testen, extrahierte Gasser alle Bestandteile des Korpus einzeln nacheinander (vgl. Gasser 2004:34). Auf diese Art und Weise konnte das Programm seine statistischen Berechnungen nicht auf den gesamten Korpus anwenden, sondern nur auf die einzelnen Dokumente.

Zerfaß hingegen beklagt lange Verarbeitungszeiten aufgrund der großen Datenmengen (vgl. 2006:24) – Fleischmann berichtet, dass große Korpora in manchen Fällen aufgrund der langen Laufzeit in kleinere Teile geteilt werden, oder auch aus strategischen Gründen geteilt werden können, z.B. um thematische Gruppierungen zu ermöglichen (vgl. Fleischmann 2006: 52).

Die Länge des Korpus bzw. verarbeiteten Textes sollte auch bei der Parametereinstellung beachtet werden, längere Texte erfordern z.B. höhere Mindestfrequenzen. Darüber hinaus kann die inhaltliche Orientierung des Textes unterschiedliche Parameter erfordern: Witschel hat festgestellt, dass Texte, die Hintergrundwissen vermitteln mit anderen Parame-

tern hochwertiger extrahiert werden können als Texte, die bereits auf umfassendem Hintergrundwissen aufbauen – letztere enthalten mehr potentielle Termini mit besonders niedrigen Frequenzen, erfordern also eine besonders niedrige Mindestfrequenz (vgl. 2005:111f.).

5.1.5.2 Testkollektionen

Die Problematik der Festlegung von Ergebnissen wurde in dieser Arbeit bereits mehrfach angesprochen. Vivaldi & Rodríguez beschäftigten sich mit der Frage, wie potentielle Termini bestimmt werden können, die als Benchmark gelten sollen und erarbeiteten folgenden Vorschlag: drei unabhängige Experten extrahieren manuell Termini aus dem entsprechenden Text bzw. Korpus, und alle Termini werden gemeinsam als Testkollektion angesehen. Es bedarf keinerlei Übereinstimmungen zwischen den Experten, um einen Terminus als gültig zu erklären – dadurch entsteht eine weit gefasste Testkollektion, die unterschiedliche Meinungen und Sichtweisen mit einbezieht (vgl. Vivaldi & Rodríguez 2007:237f.).

Eine ähnliche Idee verfolgt Thurmair: als Termini sollen all jene Token bzw. Chunks gelten, die semantische Einheiten bilden und für irgendeine terminologische bzw. translatorische Anwendung als potentielle Termini gelten könnten – als Noise sollen in diesem Verständnis nur jene Token bzw. Chunks bezeichnet werden, die für keinen denkbaren Zweck als potentielle Termini angesehen werden könnten (vgl. Thurmair 2003:175).

Kageura *et al.* rufen im Umgang mit Testkollektionen zu Vorsicht bezüglich der Gültigkeit auf: sie betonen, dass ein Abgleich der Identifikationsergebnisse mit ihrer Testkollektion nicht dahingehen interpretiert werden kann, welche Methode bzw. welches TEP besser ist (vgl. 2001:162).

Unterschiede gibt es auch bei der Qualifikation der Verantwortlichen für die manuelle Extraktion. In terminologischen Arbeiten wird die Extraktion von den Terminologen selbst durchgeführt (vgl. Fleischmann 2006:58). Computerlinguisten tendieren dazu, Fachleute mit dieser Aufgabe zu betrauen (vgl. Vivaldi & Rodríguez 2007:237), bzw. Fachgebiete auszuwählen, für die sie sich selbst als Experten qualifiziert sehen (vgl. Witschel 2005:72f.). Für ein größeres Maß an Zuverlässigkeit im Rahmen der manuellen Extraktion, werden in einigen Fällen Wörterbücher bzw. Lexika verwendet, um die Grenzen der Gemeinsprache zu definieren (vgl. Witschel 2005:72f.).

5.1.5.3 Vorverarbeitung

Im Bereich der Vorverarbeitung können zwei Themen identifiziert werden, die Konvertierung unterschiedlicher Formate und die Annotation der verwendeten Korpora. Bereitgestelltes Material ist oft in sehr unterschiedlichen Formaten gespeichert, die zunächst konvertiert werden müssen, um die Extraktion zu ermöglichen. Darüber hinaus enthalten einige Doku-

mententypen Zusatzinformationen innerhalb des Fließtextes in Form von sogenannten Tags. Wie bereits in Kapitel 2.6.2 festgestellt können einige TEP diese Formate direkt extrahieren, andere können durch Tags in ihrem Extraktionsvorgang behindert werden (vgl. Gasser 2004: 21-24). Zerfaß stellte ebenfalls Formatinkompatibilitäten fest, die einen erheblichen Mehraufwand im Vorfeld einer Evaluation bedeuten (vgl. Zerfaß 2006:24).

Im Rahmen des NTCIR-Workshops wurde das Testmaterial in annotierter und nichtannotierter Form bereitgestellt, um den einzelnen Teams die Möglichkeit zu geben, über eine gemeinsame Basis der Vorverarbeitung zu verfügen und nur die Identifikationsmethoden zu vergleichen, bzw. um besonders effiziente Vorverarbeitungssysteme verwenden zu können, die zur Qualität des Identifikationsergebnisses beitragen könnten (vgl. Kageura *et al.* 2001: 158f.). Vivaldi & Rodríguez würden die Verwendung von allgemein anerkanntem annotierten Testmaterial für Evaluationszwecke begrüßen (vgl. 2007:228 & 244).

5.1.6 Vergleichbarkeit & Standardisierung

Die Motivation hinter zahlreichen Evaluationen von TEP ist die Suche nach dem besseren, geeigneteren Programm – es handelt sich um Adäquatheitsevaluationen. Jedoch wird von vielen Evaluatoren festgestellt, dass die Vergleichbarkeit zwischen den Programmen aufgrund vielfältiger Faktoren kaum gewährleistet ist. Zwei dieser Faktoren sind die unterschiedlichen Identifikationsmethoden auf denen die Programm aufbauen, sowie die verschiedenen Einstellmöglichkeiten, die keine „einheitlichen“ Einstellungen zulassen (vgl. Zerfaß 2006:24) und somit keine zuverlässigen Ergebnisse liefern können.

Auch die Spezialisierung eines Programms auf bestimmte Methoden erschwert Evaluationen: Wanhoff evaluiert ein TEP, dass besonders stark auf kundenspezifische Stoppwortlisten angewiesen ist. Zum Zwecke der Repräsentativität und um die Zuverlässigkeit zu erhöhen werden diese Stoppwortlisten im Rahmen der Evaluation nicht verwendet, dadurch ist das Programm benachteiligt, da es seinen Vorteil nicht nutzen kann – die Gültigkeit der Ergebnisse ist dementsprechend eingeschränkt (vgl. Wanhoff 2007:25 & 40). Eine Übertragung der Stoppwortliste auf andere Programme kann in manchen Fällen eine Lösung sein, Zerfaß stellt jedoch fest, dass es sehr schwer ist, in verschiedenen Programmen sicherzustellen, dass dieselben Wörter in der Stoppwortliste enthalten sind – erneut ist die Zuverlässigkeit eingeschränkt (vgl. Zerfaß 2006:24). Auch Witschels TEP darf seine Stärken nicht zeigen: das Programm basiert auf Interaktivität mit dem Anwender – um die Objektivität zu erhalten werden alle Parameter in den Standardeinstellungen beibehalten. Als Lösungsansatz variiert Witschel die Parameter in einer nachfolgenden Analyse der einzelnen Komponenten und kann somit die Gültigkeit der Ergebnisse sicherstellen (vgl. Witschel 2005:73). Mit dem Problem der Anwenderbeteiligung hat sich auch King auseinandergesetzt: sie stellt fest, dass bei

Programmen, die stark von der Interaktion mit dem Anwender abhängig sind und ohne den Anwender nicht funktionieren, Wege gesucht werden müssen, um die Zuverlässigkeit einer Evaluation zu erhalten und sie nicht dem Talent des jeweiligen Anwenders zu überlassen (vgl. 2005:55f.).

Um eine erhöhte Vergleichbarkeit zu gewährleisten korrigiert Wanhoff im Korpus enthaltene Rechtschreib- und Formatierungsfehler vor der Extraktion – dadurch wird allerdings die Realitätsnähe des Testmaterials gestört (vgl. Wanhoff 2007:25) und es gibt keine Möglichkeit zu vergleichen, wie unterschiedliche TEP mit Texten mit höherer Realitätsnähe umgegangen wären.

Der Wunsch nach größerer Vergleichbarkeit und Standardisierung ist groß, jedoch ist es nicht einfach „goldene Standards" für die Evaluation von TEP zu bestimmen (vgl. Vivaldi & Rodríguez 2007:227). Vivaldi & Rodríguez schlagen nichtsdestotrotz vor, standardisierte Grundlagen zum Korpusdesign, Maße und Instrumente zu entwickeln – diese jedoch hoch parametrisierbar zu gestalten, um eine Anpassung an die unterschiedlichen Situationen und Bedingungen zu ermöglichen (vgl. 2007:244).

Auch L'Homme *et al.* betonen, dass für komparative Studien eine detailliert ausgearbeitete Methodologie notwendig ist, die auch das Testmaterial und seine Einflussfaktoren beachtet sowie mehrere Kriterien miteinander kombiniert und an die jeweiligen Bedürfnisse des Anwenders angepasst wird – es reicht nicht aus lediglich irgendeinen Text zu extrahieren und im Anschluss die gefundenen potentiellen Termini zu zählen (vgl. L'Homme *et al.* 1996: 309f.).

5.2 Quantitative Evaluation des Identifikationsergebnisses

Im vorigen Kapitel wurde ein Einblick in bis dato durchgeführte Evaluationen von TEP, ihre Methodologie, das verwendete Testmaterial sowie die Vergleichbarkeit unterschiedlicher Programme und Evaluationen gegeben. Als Schwerpunkt der Evaluationskriterien, Methoden und auch als bedeutendes Thema im Bezug zum Testmaterial ist die quantitative Evaluation des Identifikationsergebnisses als wichtigstes und häufigstes Merkmal im Bereich der Evaluation von TEP an zahlreichen Stellen behandelt worden. In diesem Kapitel sollen einige wichtige Aspekte zur quantitativen Evaluation des Identifikationsergebnisses zusammengefasst und diskutiert werden.

Zunächst soll noch ein kleiner Rückblick erfolgen: Die einzelnen Werte Recall, Precision, Noise und Silence sowie der F-Wert wurden in Kapitel 2.4.1 eingeführt. Die Auswirkungen der einzelnen Identifikationsmethoden auf diese Werte wurden unter dem Unterpunkt Forschungsmeinungen in den Kapiteln 2.4.3 und 2.4.4 beschrieben, in Kapitel 3.3.1 wurde im Rahmen der Beschreibung bekannter Probleme von TEP auf die Qualität des Identifikations-

ergebnisses eingegangen. In Kapitel 4.4.3.2 wurden Baselines und Benchmarks als Bestandteile von Evaluationsmethoden des EAGLES-Rahmen eingeführt, und in Kapitel 4.4.4.2 wurde näher auf Benchmarks als systematischen Testtyp eingegangen. Testmaterial, insbesondere Testkollektionen im Sinne des EAGLES-Evaluationsrahmen wurden in Kapitel 4.4.6 behandelt. In allen Unterkapiteln des Kapitel 5.1 spielt die Qualität des Identifikationsergebnisses eine bedeutende Rolle. In Folge sollen einige Unzulänglichkeiten bei der Verwendung der Werte zur quantitativen Evaluation des Identifikationsergebnisses aufgezeigt werden.

Mangelhafte Dokumentation

In zahlreichen Publikationen werden direkt nach einer kurzen Beschreibung der Evaluation die ersten Zahlen präsentiert. Oft ist nicht klar, auf welcher Grundlage diese Werte entstanden sind. Es fehlen Daten z.B. zur Korpusgröße (möglichst detaillierter als Seitenzahlen), zu den enthaltenen Textarten und zur Themenverteilung, zur Verarbeitung des Korpus (in Teilen oder vollständig), zur Vorverarbeitung des Korpus, zur Einstellung der Parameter, zur Verwendung zusätzlicher linguistischer Daten und ihrer Beschaffenheit (z.B. Stoppwortlisten), zur Validierung und den Kriterien anhand derer validiert wurde, zur Erstellung der Testkollektion (von wem die vorbereiteten Ergebnisse ausgewählt wurden, anhand welcher Kriterien die manuelle Extraktion durchgeführt wurde, ob hierzu Wörterbücher oder Lexika zur Abgrenzung der Gemeinsprache verwendet wurden, ob die manuelle Extraktion vor oder nach der computergestützten durchgeführt wurde), zum Abgleich der Testkollektion mit dem Identifikationsergebnis (manuell, mithilfe von Tabellenkalkulationen oder anderer Software), zur Wertung von Wortwiederholungen und Tippfehlern, zu den Auswirkungen eventueller Grundformenbildungen, etc.

In zahlreichen Fällen wurden diese Informationen vermutlich intern dokumentiert und sind nur nicht in die Publikation übernommen worden. Es sind auch nicht alle angeführten Daten im Rahmen jeder Evaluation relevant, jedoch erlaubt die insgesamt mangelhafte Dokumentation kaum eine Beurteilung von nicht selbst durchgeführten Evaluationen oder Rückschlüsse auf die verwendete Methodologie für eine eventuelle Übernahme der Methoden für eine eigene Evaluation. Aufgrund der fehlenden Informationen erscheint ein Vergleich mit weiteren Daten einer anderen Evaluation ebenfalls nur sehr eingeschränkt und oberflächlich möglich. Gute Beispiele für ausführliche Dokumentation sind die Arbeiten von Witschel und Fleischmann, die zahlreiche Details zur Methodologie, zum Korpus und zur Berechnung der Recall- und Precision-Werte enthalten (vgl. Witschel 2005:71-75, Fleischmann 2006:69f., 95-119).

Divergierendes Quantitätsverständnis

Es gibt selbstverständlich keine einheitlichen Vorstellungen davon, ab welchem Wert von „viel Noise“ bzw. „zu viel Noise“ gesprochen wird. Für Cabré *et al.* produziert z.B. das TEP Termino mit 28% zu viel Noise (vgl. 2001:77). Vermutlich ist diese Beurteilung dadurch entstanden, dass alle anderen Systeme im Vergleich, die Noise bzw. Precision-Werte angegeben haben, um 5-26% unter diesem Wert lagen (vgl. Cabré *et al.* 2001:82). Die verglichenen Werte wurden anderen Publikationen entnommen, es handelt sich nicht um eine gemeinsame vergleichende Evaluation.

Zum Vergleich: Bei Fleischmann lagen die reinen Noise-Werte der vier evaluierten TEP z.B. direkt nach der Identifikation bei 43,45-55,60% und nach der Verwendung der jeweiligen Filterungsfunktionen bei 26,47-46,19%. In fast allen Fällen überstiegen sie somit den bei Cabré *et al.* als „zu viel“ bezeichneten Prozentsatz von 28%, von dem nicht bekannt ist, wie er entstanden ist. Dieses Beispiel zeigt, dass keine grundsätzliche Vergleichbarkeit gegeben ist. Nichtdestotrotz werden – auch von den Evaluatoren selbst – oft Formulierungen wie „zu viel“ oder „ein gewisses Maß“ bzw. „relativ hohe Noise-Werte“ verwendet, ohne eine Begründung anzufügen, die erklärt, auf welcher Grundlage diese Wertungen entstanden sind. Aus diesem Grund finden sich auch in dieser Arbeit in Kapitel 3.3.1 Formulierungen dieser Art, da sie aus den konsultierten Publikationen übernommen wurden (vgl. z.B. Cabré *et al.* 2001:83).

Zur Verdeutlichung der Relativität soll hier noch ein weiteres Beispiel dienen: Gasser gibt die Anzahl der Termkandidaten nach der Identifikation und dem Abgleich mit Ausschlussdatenbanken mit 17.272 an, 447 wurden als Termini bestätigt. Es wird kein Noise-Wert berechnet, aber der Prozentsatz der nicht bestätigten Termini läge in diesem Fall bei über 97%. Dennoch zeigt sich die Autorin zufrieden und bestätigt der computergestützten Extraktion eine kürzere Bearbeitungszeit als der manuellen Extraktion (vgl. Gasser 2004: 36f.).

Falls keine zuverlässigen Methoden für die Evaluation aller am Vergleich beteiligten TEP verwendet wurden, sollten keine vergleichenden Schlussfolgerungen gezogen werden.

Problematik Testkollektionen

Die für die Berechnung der Silence- bzw. Recall-Werte notwendigen Testkollektionen führen durch den erforderlichen Arbeitsaufwand meist zu einer Reduktion des Korpus, und reduzierte Korpora zu einer reduzierten Repräsentativität der Ergebnisse (vgl. Kapitel 5.1.5.1). Testkollektionen werden oft in Eigenarbeit des Evaluators ohne die Verwendung dokumentierter Kriterien erstellt und sind somit in manchen Fällen nicht zuverlässig – Methoden zur Bestimmung von zuverlässigen Testkollektionen wurden in Kapitel 5.1.5.2 besprochen.

Offen bleibt die Frage, ob der jeweilige Abgleich von Ergebnissen mit Testkollektionen und der darauffolgende Rückschluss auf die Güte der jeweiligen TEP gültig ist (vgl. Kageura *et al.* 2001:162). Das Identifikationsergebnis eines TEP hängt von zahlreichen Faktoren ab (Korpusgröße, -beschaffenheit, Sprache, Vorverarbeitung des Korpus, Identifikationsmethoden, ausgewählte Parameter). Die Beschaffenheit der Testkollektion hängt ebenfalls von zahlreichen Faktoren ab (extrahierende Person, Qualifikation, Vorgaben zu Validierungskriterien, Rahmenbedingungen [Verwendung von Lexika], Arbeitsumgebung [manuell, mithilfe eines anderen Programms], Zweck der Extraktion). Sofern nicht möglichst viele dieser Faktoren als Bestandteile der Methodologie dokumentiert werden, muss die Repräsentativität der Evaluation auf die ursprüngliche Evaluation beschränkt bleiben und es können nur sehr oberflächliche Rückschlüsse auf die universale Befähigung des TEP zur Extraktion von Termini gezogen werden (vgl. EAGLES 1995:36f., L'Homme 1996:303).

Eine Möglichkeit zur Nutzung dieser beschränkten Repräsentativität ist die Analyse eines bestimmten Phänomens anhand der Ergebnisse, z.B. nutzt Wanhoff das Identifikationsergebnis zur Analyse der Wortbildungsmuster, die das TEP nicht extrahieren kann (vgl. 2007: 71).

Alternative Kriterien

Die Verwendung von Recall, Precision und F-Wert wurde aus dem Bereich des Information Retrieval (IR) übernommen (vgl. Ahmad & Rogers 2001:747f.). Bei dieser Entscheidung wurde vielfach außer Acht gelassen, dass die Extraktion in diesem Bereich anderen Anforderungen unterliegt als in der Terminologieextraktion: für IR-Zwecke sollen die wichtigsten Termini bzw. Wörter und Wortfolgen extrahiert werden, für terminologische und translatorische Zwecke sollen alle relevanten Termini bzw. Wörter und Wortfolgen extrahiert werden – und relevant hängt in diesem Fall stark vom Anwender bzw. vom Zweck der Extraktion ab (vgl. Thurmair 2003:174, Cabré *et al.* 2001:54). Thurmair stellt fest, dass zahlreiche Evaluatoren diesen grundsätzlichen Unterschied nicht beachten, und die Evaluationsergebnisse dadurch an Wert verlieren (vgl. 2003:175).

Auch im Bereich des IR wurde inzwischen festgestellt, dass absolute Recall-Werte schwer zu definieren sind, wenn nicht eindeutig festzustellen ist, was gefunden hätte werden können – deshalb wurde z.B. die Vorstellung des Recalls als absoluten Wert durch einen relativen Vergleichswert ersetzt (vgl. Sparck-Jones & Galliers 1995:42). Spark-Jones & Galliers betonen, dass die direkte Übernahme der IR-Erfahrungen in einige weitere NLP-Bereiche positiv gesehen werden kann. Es darf jedoch nicht vergessen werden, dass es sich um unterschiedliche Bereiche handelt und eventuell weitere Entwicklungen von eigenen Kriterien

notwendig sind (vgl. 1995:29), die jedoch ohne Zweifel sehr schwer zu definieren sind (vgl. 1995:154).

Viele der weiteren verwendeten Kriterien beziehen sich ebenfalls auf direkte oder indirekte Art auf die Qualität des Identifikationsergebnisses. L'Homme *et al.* identifizieren zahlreiche Kriterien zur Vorevaluation, ein Großteil dieser Kriterien ist eng mit dem Identifikationsergebnis verbunden: Identifikationsmethode, Art der identifizierten Termini, zahlreiche identifikationsbezogene Parameter – lediglich die Sprache, der Anwendungszweck, die Eigenständigkeit des Programms und die Möglichkeiten zu Nachbearbeitung weisen keinen direkten Bezug zur Qualität des Identifikationsergebnisses auf und können somit unabhängig von Precision- und Recall-Werten bewertet werden (vgl. L'Homme *et al.* 1996: 294-309). Bei Cabré *et al.* sind zwei der erweiterten Vergleichskriterien, Strategien der Tokenisierung und Filterung, mit dem Identifikationsprozess verbunden und können nur durch ihn bewertet werden. Die beiden weiteren Kriterien können auch unabhängig beurteilt werden: es handelt sich um die Abhängigkeit von linguistischem Material sowie die Hilfestellung bei der Gruppierung von potentiellen Termini im Rahmen der Nachbearbeitung (vgl. Cabré *et al.* 2001:79ff.).

In einigen Fällen werden zu Beginn der Evaluation zahlreiche Kriterien genannt, die im Rahmen der Evaluation angewendet werden sollen. Leider wird die Evaluation im Verlauf aber beständig reduziert, und im Fazit zählt nur mehr die quantitative Evaluation des Identifikationsergebnisses. Bei Cabré *et al.* wird ein Vergleich unter Berücksichtigung von sechs verschiedenen Aspekten durchgeführt, als Resultat werden allerdings nur die Recall- und Precision-Werte bezeichnet und in den Schlussfolgerungen spielen fast nur quantitative Aspekte eine Rolle (vgl. 2001:79-83).

Zu Beginn von Fleischmanns Arbeit wird eine Vielzahl an Evaluationskriterien genannt, die bei einer Evaluation beachtet werden sollten (vgl. Fleischmann 2006:45). Jedoch wird bei der Definition der Anforderungen, die wiederum erfreulich umfassend ausfällt (Zeitersparnis, Kostenersparnis, verteilter Workflow, Outsourcing, Nachbearbeitung, Bedienerfreundlichkeit, Formatunterstützung), von einem eingeschränkten Qualitätsverständnis ausgegangen: die Qualität des Programms wird mit der Qualität der potentiellen Termini gleichgesetzt (vgl. 2006: 39ff.). Auf einige der anfangs genannten Kriterien wird im Rahmen der Evaluationsbeschreibung eingegangen (vgl. 2006: 98-109), in der Auswertung der Ergebnisse wird allerdings nur noch das Identifikationsergebnis im Rahmen von statistischen Berechnungen und Kostenfaktoren betrachtet (vgl. 2006:109-124). Dies kann selbstverständlich an den Anforderungen des Auftraggebers liegen, der sich in diesem Fall eine Analyse der Wirtschaftlichkeit unter der Berücksichtigung der Qualität gewünscht hatte (vgl. Fleischmann 2006:69f.). Eine Aufgabe des Evaluators bestand in diesem Fall darin, Qualität auf die An-

forderungen des Auftraggebers herunterzubrechen: das Endergebnis war eine starke Betonung der Funktionalität (vor allem der Richtigkeit und Angemessenheit) sowie der Effizienz (insbesondere des Zeitverhaltens).

Einen anderen Weg wählte Thurmair, der sich z.B. auf das Kriterium Benutzbarkeit konzentrierte und drei Formen der Benutzbarkeit formulierte, die ausschlaggebend für die Qualität eines TEP sein sollen: 1. sollte der Silence-Wert möglichst gering sein, da ein TEP nur dann etwas „nützt", wenn keine manuelle Extraktion mehr notwendig ist; 2. sollte die Validierungszeit möglichst kurz sein (hiermit sind sowohl geringe Noise-Werte als auch Unterstützungsmechanismen in der Validierungsphase gemeint); 3. sollte das Ergebnis möglichst vielfältig genutzt werden, zahlreiche Exportformate sollten unterstützt und Zusatzinformationen angeboten werden (vgl. Thurmair 2003:175f).

5.3 Ganzheitliche Evaluationsmethodologien

Die Arbeiten von Lieske (2002) und Saß (2004) wurden anhand des EAGLES-Evaluationsrahmens entwickelt. Daher soll hier die Möglichkeit genutzt werden, sie als Beispiele für eine Anwendung des Rahmens vorzustellen.

5.3.1 Methodologische Grundlagen

Lieske verfolgt die Idee des Consumer Report Paradigms nicht, die Evaluation wird anhand eines konkreten Auftrags durchgeführt (vgl. 2002:114). Saß hat allem Anschein nach aus eigener Motivation heraus die Evaluation durchgeführt, sie verweist auf keinen konkreten Auftrag, sondern stellt fest, dass die Evaluation aus Übersetzersicht durchgeführt wird. Jedoch werden keine Evaluationsziele oder Anforderungen der Evaluation genannt, an denen die Evaluationsmethodologie ausgerichtet wurde. Das von ihr entwickelte Vergleichsmodell lässt vermuten, dass die Evaluation einen universalen Charakter haben sollte (vgl. Saß 2004: 29-36).

Der EAGLES-Evaluationsrahmen

Die EAGLES-Grundlagen sind in Lieskes Arbeit deutlich zu erkennen: nach einer Einführung in TEP wird ein Ausschnitt des Projektplans präsentiert, der das Evaluationsziel und die zu evaluierenden Funktionen enthält (vgl. 2002:114f.). Im Anschluss wird das Qualitätsmodell der ISO-IEC Norm 9126 vorgestellt und Ausschnitte einer Liste wesentlicher informativer Merkmale präsentiert. Diese Liste enthält neben den Merkmalen auch die ihnen zugrunde liegenden Anforderungen sowie die Methoden bzw. Maße, die den Merkmalen zugeordnet wurden (vgl. Lieske 2002:115ff.). Es folgen allgemeine Angaben zum erfor-

derlichen Testmaterial sowie zum Evaluationsablauf unter Zuordnung der jeweiligen Verantwortlichen (vgl. 2002:117f.).

Auch in Lieskes Evaluation ist die Analyse des Identifikationsergebnisses von besonderer Bedeutung, zu diesem Zweck befassen sich zwei Kapitel mit der Beschreibung der Evaluationsmethoden des Identifikationsergebnisses: Anforderungen an die Identifikationsfunktion, Klassifikationsschema als objektive Entscheidungsgrundlage bei der Validierung der potentiellen Termini, Entwicklung eines Tools zur möglichst effizienten und genauen Validierung, Festlegung der wesentlichen informativen Merkmale aus Sicht der Anwender & Käufer (vgl. Lieske 2002: 118-123). Abschließend werden der Extraktions- und Validierungsprozess beschrieben und einige Ergebnisse präsentiert, in denen die statistischen Ergebnisse mit für eine Interpretation durch den Leser hilfreichen Informationen ergänzt werden (vgl. Lieske 2002:123-128).

In der Arbeit von Saß erscheinen die EAGLES-Grundlagen nicht derart offensichtlich, wie es bei Lieske der Fall ist. Der Verweis auf die EAGLES-Gruppe (vgl. Saß 2004:17) und die von ihr erstellte Kriterienchecklisteste sind die sichtbarsten Zeichen im Sinne des EAGLES-Evaluationsrahmens. Die Entstehung der Kriterienchecklisteste ist nicht dokumentiert, allerdings sind sowohl die vorbereitete Liste als auch die ausgefüllte Liste vollständig im Anhang enthalten (vgl. Saß 2004:120-130).

Die anwenderbezogene Orientierung wird bei der Beschreibung der unterschiedlichen Ausgangssituationen bei monolingualen bzw. bilingualen Extraktionen und den entsprechenden Aufgaben deutlich – diese Übersicht entspricht in etwa dem im EAGLES-Rahmen vorgeschlagenen Prozessmodell (vgl. Saß 2004:21). In Folge wird der Evaluationsablauf beschrieben, bezogen auf die Durchführung der einzelnen Extraktionen (vgl. Saß 2004:21f.). Das Kapitel zur Methodologie enthält Informationen zum Testmaterial und Auszüge aus einer manuell erstellten Testkollektion, sowie Hinweise auf die verwendete Literatur zur Erstellung der Kriterienchecklisteste (vgl. Saß 2004:27ff.). Im Anschluss werden die einzelnen Kriterien der Liste vorgestellt, jedoch ohne Angaben zu Maßen oder Methoden (vgl. 2004:30-36). Die im Anhang enthaltene Liste enthält diese Informationen ebenfalls nicht, zu jedem Kriterium wird eine Frage gestellt – Maße in Form von Antwortmöglichkeiten sind nicht enthalten (vgl. Saß 2004:120-122).

Der erste Teil der Evaluation besteht aus der Kommentierung der einzelnen Ergebnisse der Kriterienchecklisteste, zu einigen Kriterien (z.B. Identifikation und Grundformenbildung) werden quantitative Ergebnisse präsentiert (vgl. 2004:65-74). Der zweite Teil der Evaluation besteht auch bei Saß aus einer näheren Betrachtung der Qualität des Identifikationsergebnisses bzw. der Fehler im Rahmen der Grundformenbildung. Angesichts der durch die Autorin erstellten Testkollektion wurde keine gesonderte Validierung durchgeführt und dementspre-

chend auch keine Methodologie erarbeitet. Die Parameter wurden in mehreren Testläufen variiert, um die Ergebnismarge vom jeweils besten Silence- bis zum besten Noise-Wert abzudecken (vgl. Saß 2004:75-88).

Evaluationskriterien

Da Lieske nur Ausschnitte aus der Merkmalliste präsentiert hat, können keine Rückschlüsse auf den Umfang der verwendeten Kriterien gezogen werden. Bekannt sind lediglich folgende Merkmale: Extraktion Deutsch/Englisch sowie Sprachen mit anderem Zeichensatz, Bereitstellung von linguistischen Ressourcen, Unterstützung der wichtigsten Microsoft Word-Formate, „günstiges" Verhältnis von Noise zu brauchbaren potentiellen Termini, Zusatzinformationen zu Häufigkeit, Wortart und Fundstelle, Grundformenbildung, Exportmöglichkeiten (vgl. Lieske 2002:116f.). Dieser Ausschnitt lässt eine recht umfangreiche Definition an Merkmalen zumindest vermuten.

Saß Kriterienchecklisteist die bisher umfangreichste dokumentierte Kriteriensammlung zur computergestützten Terminologieextraktion[30], die auch tatsächlich in einer Evaluation angewandt wurde. Sie umfasst die Bereiche *Technische Beschreibung, Bedienungsoberfläche, Terminologieextraktion, Eingabe und Import, Ausgabe und Export, Validierung der Ergebnisse* und *ökonomische Aspekte*.

5.3.2 Qualität des Identifikationsergebnisses

Auch diese beiden Evaluationen widmen einen Schwerpunkt ihrer Evaluation der Qualität des Identifikationsergebnisses. Die Methodologie weist große Unterschiede auf, auch die Ausgangssituationen sind sehr unterschiedlich. Allein die personellen Ressourcen sind nicht vergleichbar: Lieskes Validierung von insgesamt 10000 potentiellen Termini wurde von 9 technischen Redakteuren bzw. Übersetzern durchgeführt (vgl. 2002:123), Saß extrahierte die Termini als Bestandteil der Testkollektion selbst aus dem insgesamt nur 4000 Wörter umfassenden Korpus (vgl. 2004:27). An dieser Stelle soll besonders betont werden, dass die unterschiedliche Methodologie auch durch die unterschiedlichen Blickwinkel beeinflusst sein kann: Saß wählte für ihre Evaluation die translatorische Sicht, Lieske einen terminologischen Ausgangspunkt.

Saß: Aus translatorischer Sicht

Saß evaluierte sowohl die mono- als auch die bilinguale Terminologieextraktion. Den Hintergrund der Evaluation bilden zwei verschiedene Szenarien: die übersetzungsvorbereitende monolinguale Extraktion zu Identifikation von im Vorfeld der Übersetzung vorzubereitender

[30] Zumindest ist der Autorin dieser Arbeit keine umfangreichere Liste bekannt.

Terminologie sowie die übersetzungsnachbereitende bilinguale Extraktion zur Gewinnung und Wiederverwendung von bilingualer Terminologie. Das Testmaterial bestand aus zwei englischen Texten in einem Gesamtumfang von 4000 Worten, die auch in deutscher Übersetzung vorlagen. Saß extrahierte manuell 144 Termini und ihre zugehörige Übersetzung (vgl. Saß 2004:27). Ausschlusskriterien waren eventuell vorhanden, dokumentiert ist lediglich, dass die Termini keinen Artikel enthalten durften (vgl. 2004:29).

Im zweiten Teil der Evaluation führte Saß „Berechnungen der Effizienz“ durch, die sich jedoch auf die Berechnung von Precision und Recall beschränken – in der EAGLES-Terminologie demnach die Funktionalität und nicht die Effizienz betreffen. Im ersten Extraktionsdurchlauf wurden in allen Programmen die Einstellungen so gewählt, dass möglichst niedrige Silence-Werte entstehen (z.B. Mindestfrequenz = 1). Zur Validierung wurden die Identifikationsergebnisse mit den Termini der Testkollektion verglichen, auch falsch lemmatisierte bzw. nicht richtig abgetrennte potentielle Termini wurden als korrekt eingestuft. Anschließend wurden alle korrekten und inkorrekten potentiellen Termini gezählt und prozentuelle Recall- und Silence-Werte berechnet. In weiteren Tests wurden die Parameter variiert, um die vergleichsweise hohen Noise-Werte (67,6 - 88,5%) zu verbessern. Es konnten nur geringfügige Verbesserungen erzielt werden (vgl. Saß 2004:75-87).

Die Identifikationsergebnisse wurden darüber hinaus auf falsche Grundformenbildung untersucht und kategorisiert, anhand der Ergebnisse wurden Fehlerquoten der Grundformenbildung berechnet (vgl. 2004:69). Hierbei ist unklar, ob die TEP alle falsch lemmatisierten Token als potentielle Termini ausgegeben haben oder weitere Fehler unentdeckt blieben und die Fehlerquoten somit ungültig wären.

Für die bilinguale TE wurden zwei TEP zur Extraktion aus TMX-Dateien verwendet, und ein TEP das zuerst monolingual extrahiert und in Folge die Übersetzung sucht. Hier wurde lediglich die Anzahl der zu 100% korrekten gefundenen Terminologiepaare angegeben, und als Zusatzinformation, an welcher Stelle der Vorschlagsliste die zugehörige Übersetzung gefunden wurde (vgl. Saß 2004:87f.).

Insgesamt sticht die von Saß verwendete Methodologie zur Bewertung der Qualität des Identifikationsergebnisses im Vergleich zu den in Kapitel 5.1.4 vorgestellten Methodologien nicht besonders hervor, und die Repräsentativität der Ergebnisse für terminologische Zwecke könnte trotz der relativ ausführlichen Dokumentation aufgrund des kleinen Korpus in Frage gestellt werden. Jedoch sollte hierbei beachtet werden, dass die Evaluation auf einen übersetzerischen Blickwinkel ausgerichtet war. Für übersetzerische Zwecke dürften die Evaluationsergebnisse ausreichend repräsentativ sein, da kurze Texte in den jeweiligen Anwendungsszenarien von Übersetzern durchaus üblich sind.

Lieske: Aus terminologischer Sicht

Lieske evaluierte die monolinguale Terminologieextraktion von fünf TEP aus vier verschiedenen Sprachen. Es wurden jeweils die Standardeinstellungen beibehalten und keine linguistischen Ressourcen verwendet. Im Vorfeld der Extraktion wurden bereits wesentliche informative Merkmale und zugehörige Maße definiert, sowie ein Klassifikationsschema als Grundlage für die Validierung der potentiellen Termini entwickelt. Dieses Klassifikationsschema ermöglichte in vier Sprachen die einheitliche Zuordnung von Charakteristika, wie z.B. *gültig*/*ungültig*, *Mehrwort*/*Einwort*, und weitere Eigenschaften wie *enthält Abkürzung* /*unvollständig*. Die Validierer wurden am Extraktionsprozess nicht beteiligt, da dieser von einem Projektmitarbeiter durchgeführt wurde, der die extrahierten potentiellen Termini aller Programme zusammenfasste und für die Validierung im Rahmen der für diese Zwecke entwickelten Validierungsumgebung vorbereitete (vgl. Lieske 2002:116-122).

Im Anschluss an die Validierung wurden statistische Berechnungen anhand der Ergebnisse durchgeführt und nach im Vorfeld festgelegten Interpretationsvorgaben mit Zusatzinformationen zum Bericht aufgearbeitet, in dem die unterschiedlichen Bedürfnisse von Anwendern und Käufern beachtet wurden. Für Anwender wurden z.B. Beispiele der Identifikationsergebnisse in den Bericht aufgenommen. Für Käufer wurden hingegen statistische Ergebnisse betrachtet, die für eine eventuelle Kaufentscheidung relevant sein könnten, z.B. TEP, die in der Anzahl der gefundenen validierten Einwortterme besonders gute Ergebnisse erzielt haben – als Anzeichen für besondere Fähigkeiten eines TEP (vgl. Lieske 2002:122-128).

Lieske hat auf die Erstellung einer Testkollektion verzichtet, und als Basiswert für Recall-Berechnungen die Gesamtzahl aller als korrekt validierten Termini genutzt (vgl. 2002: 126). Diese Methode ermöglicht Recall-Berechnungen für große Korpora, jedoch besteht die Möglichkeit, dass die Recall-Werte aller TEP durch diese Methode zu niedrig ausfallen und nicht gültig sind, falls eine bestimmte Anzahl an Termini von keinem der verwendeten TEP erkannt wurde. Die Precision-, Recall- und F-Wert-Berechnungen wurden sprachspezifisch durchgeführt, so wurde unter anderem festgestellt, dass die Fähigkeit der TEP, vom Anwender gewünschte Termini zu identifizieren, je nach Sprache variieren kann (vgl. Lieske 2002: 125-128).

Die Methodologie von Lieske zeichnet sich durch die Bemühungen aus, möglichst objektive, formalisierte Methoden und Maße zu finden, die möglichst genau dokumentiert sind und durch den damit verbundenen Versuch, den subjektiven Einfluss durch menschliche Interaktion zu reduzieren.

5.3.3 Vergleichbarkeit & Standardisierung

Der Aufwand für die Durchführung einer derart formalisierten und umfangreichen Evaluation ist jedoch sehr hoch. Dementsprechend beklagt Lieske die mangelnde Standardisierung der Import- und Exportformate, nicht verfügbares vorbereitetes Testmaterial (insbesondere Testkollektionen) und mangelhafte Integrationsmöglichkeiten der TEP in Softwareumgebungen (kein standardisiertes API), die die Durchführung von Vergleichen erschweren (vgl. 2002: 118 & 128ff.).

Im Fazit äußert auch Saß den Wunsch nach einem standardisierten Evaluationsmodell und definierten Toleranzgrenzen für die Akzeptanz von potentiellen Termini (vgl. Saß 2004: 89). Ähnliche Wünsche wurden bereits im Kapitel 5.1.6 als Fazit aus anderen bisherigen Evaluationen von TEP gezogen.

5.4 Ideen zur Entwicklung eines parametrisierbaren Evaluationsmodells

Der Wunsch nach standardisierten Evaluationsmethodologien bzw. einzelnen standardisierten Merkmalen, Methoden oder Testmaterialien wird von zahlreichen Evaluatoren geäußert (siehe Kap. 5.1.6 und 5.3.2). Jedoch sind die Ausgangspunkte und Wünsche so divers wie die Anwendungszwecke der computergestützten Terminologieextraktion (siehe Kap. 1.1); und die Standardisierung kann – wie bereits in Kapitel 4.1.1.4 angesprochen – nicht anhand von starren Vorgaben erfolgen. Eine Möglichkeit zur Vereinheitlichung würde ein parametrisierbares Modell gemäß den Grundlagen des EAGLES-Evaluationsrahmens bieten.

Ein Beispiel für eine komplexe Umsetzung eines parametrisierbaren Evaluationsmodells ist das *Framework for the Evaluation of Machine Translation in ISLE* (Femti), das eine webbasierte Methodologieerstellung für Evaluationen von Programmen zur maschinellen Übersetzung ermöglicht (vgl. Estrella *et al.* 2009). Zunächst können der Kontext der zu testenden Anwendung beschrieben und eigene Anforderungen festgelegt werden. Das Framework schlägt dem Anwender in Folge geeignete Qualitätskriterien, Kriterienbereiche und einzelne Kriterien vor, von denen die gewünschten durch den Evaluator ausgewählt werden. Das Programm bietet Testmethoden zur Auswahl, die geringfügig modifiziert werden können. Als Endergebnis erhält der Evaluator einen detaillierten Evaluationsplan mit Erklärungen zur Bedeutung der einzelnen Kriterien und zur Durchführung der einzelnen Tests (vgl. Estrella *et al.* 2009:6-10).

Die Erarbeitung eines derartig vollständigen parametrisierbaren Modells für die Evaluationsmethodologie von Terminologieextraktionsprogrammen ist im Rahmen dieser Masterarbeit leider nicht möglich. Bereits die systematische Erfassung von Daten für die ausführliche Betrachtung verschiedener Anwenderprofile könnte in einer eigenen Masterarbeit erfolgen. Aus diesem Grund sollen in diesem Kapitel lediglich Ideen aus der Vielzahl an be-

trachteten Arbeiten zusammengetragen werden und grundsätzliche Tendenzen für die Entwicklung eines parametrisierbaren Evaluationsmodells beschrieben werden.

Die einzelnen im EAGLES-Evaluationsrahmen vorgesehenen Schritte zur Erarbeitung einer Methodologie für eine anwenderorientierte Adäquatheitsevaluation werden der Reihe nach betrachtet: zu jedem einzelnen Schritt werden Gedanken zur weiteren Entwicklung gesammelt und soweit möglich Ideen formuliert. Diese Ideen sollen als erster Ansatz zur Entwicklung eines parametrisierbaren Modells gesehen und lediglich als Entwurf aufgefasst werden, der sich im Entwicklungsstadium befindet – es besteht keinerlei Anspruch auf Vollständigkeit.

Das angestrebte parametrisierbare Modell sollte sowohl für allgemeine als auch für spezifische Adäquatheitsevaluationen angewendet werden können, auf spezifische Evaluationen wird aus diesem Grund jeweils gesondert eingegangen (für die Unterscheidung von allgemeinen und spezifischen Evaluationen siehe auch Kap. 4.3.1.1).

Im Anschluss an dieses Unterkapitel werden die beiden nach Meinung der Autorin bestgeeigneten Testtypen zur Evaluation von TEP (Kriterienüberprüfung und Benchmarktests) erneut thematisiert, um auch in diesem Bereich einige Ideen zu sammeln bzw. konkrete Entwürfe zu entwickeln (Kap. 5.5).

5.4.1 Ziele & Ausgangssituation

Als ersten Schritt im Zuge der Planung einer Evaluation müssen die Ziele der jeweiligen Evaluation festgelegt werden. In Kapitel 4.3.1 wurden einige Überlegungen zur Analyse der Evaluationssituation getroffen. Anhand dieser Überlegungen und der unterschiedlichen Ziele der Anwender von Terminologieextraktionsprogrammen (siehe Kap. 1.1) wurden folgende Fragen entwickelt, die zur Definition der Ziele beantwortet werden sollten:

- Für welche Zielgruppe soll evaluiert werden? (Alle potentiellen Anwender [Anwender/Käufer], Terminologen [Anwender/Käufer], Translatoren [Anwender/Käufer], Projektmanager, Lexikographen, Sonstige)
- Welche Programme sollen evaluiert werden? (Alle derzeit am Markt erhältlichen/eine bestimmte Art von derzeit am Markt erhältlichen [für eine bestimmte Sprache, bestimmtes Identifikationsverfahren, monolinguale/bilinguale]/alle unentgeltlich/bis zu einem bestimmten Preis verfügbaren Programme/alle eigenständigen TEP/alle TEP mit besonderen Eigenschaften [z.B. Konsistenzprüfunktionen/Kompatibilität mit Datenbank]/Sonstige)
- Zu welchem Zweck erfolgt die Evaluation? (Marktüberblick/Aufzeigen neuer Funktionen/Vergleich zu manueller Extraktion/Sonstiges)

- Welche Funktionen der TEP sollen evaluiert werden? (sämtliche/Vorbereitung/ Identifikation/Nachbearbeitung/Export/Validierung/Zusatzfunktionen)
- Im Hinblick auf welchen Anwendungszweck soll die Evaluation erfolgen? (sämtliche Anwendungsmöglichkeiten/Terminologiearbeit [systematische/punktuelle, begriffsorientierte/benennungsorientierte]/Übersetzerische Tätigkeiten [monolinguale/bilinguale]/Lexikographie/Sonstiges)
- Welche Qualitätskriterien sollen besonders priorisiert werden? (Funktionalität/Zuverlässigkeit/Benutzbarkeit/Effizienz/Änderbarkeit/Übertragbarkeit)

Im Rahmen einer spezifischen Evaluation können bereits weitere Ziele definiert werden, die in einer allgemeinen Evaluation im Rahmen der Anwenderprofile behandelt werden. Für spezifische Evaluationen ist die Erstellung von Anwenderprofilen nicht zwingend nötig, da die Bedürfnisse der potentiellen zukünftigen Anwender direkt analysiert werden können, und zwar anhand folgender Fragen:

- Welche Ziele werden mit dem Erwerb des TEP verfolgt? (Anwender/Käufer)
- Welcher Nutzen wird vom TEP erwartet? (Anwender/Käufer)
- Wer wird das Programm benutzen?

Auch die Ausgangssituation kann im Rahmen von spezifischen Evaluationen direkt analysiert werden. Hierfür sollten die folgenden Fragen beantwortet werden:

- Wie sieht das jeweilige Setup aus? (welche Hardware ist vorhanden, auf wie vielen Computern soll das TEP installiert werden, in welche Softwareumgebung soll das TEP integriert werden, welche Schnittstellen stehen zu internen Datenbanken/CMS zur Verfügung)
- Wie sieht die jeweilige Umgebung aus? (wie viele zukünftige Anwender, welche Kenntnisse, wie sieht der Workflow aus)

Anhand dieser Fragen, die entsprechend der Evaluationssituation noch erweitert werden können, kann bereits eine Analyse der Ausgangssituation erfolgen, die für die weitere Evaluationsplanung möglichst dokumentiert werden sollte. Eine ausführliche Analyse zu Beginn der Planung, die auch bereits nicht zu beachtende Faktoren und Bereiche ausschließt, ermöglicht eine zielgerichtete Evaluation und ist für eine erfolgreiche Durchführung unentbehrlich (vgl. Lieske 2002:114). Auf die Analyse der Ziele und der Ausgangssituation erfolgt bei der Planung von allgemeinen Evaluationen die Erstellung der Anwenderprofile. Diese Aufgabe soll im nächsten Unterkapitel betrachtet werden.

5.4.2 Anwenderprofile

Eine Anwendercharakterisierung ist eine umfangreiche Aufgabe, die im bestmöglichen Falle durch eine Datenerhebung unter potentiellen Anwendern erfolgt (vgl. EAGLES 1995:101 &

161f.). Die in der Online-Umfrage (*t-survey*) von Zielinski & Ramírez-Safar (2005 & 2006) erhobenen Daten könnten eventuell für diesen Zweck verwendet werden, sind aber leider nicht in vollständigem Umfang verfügbar. Dementsprechend können hier in Ermangelung umfassender Daten lediglich einige Ideen zur Anwendercharakterisierung gesammelt werden.

Im Rahmen der Zieldefinition wurde die Zielgruppe der Evaluation festgelegt. Die Anwenderprofile können auf diese Zielgruppe beschränkt werden, falls die beiden Personengruppen übereinstimmen. Dies ist z.B. nicht der Fall, wenn die Evaluation an Käufer der Programme gerichtet werden soll, jedoch die Bedürfnisse der Anwender in die Evaluation mit einbezogen werden.

Es könnten z.B. folgende Charakteristika bei der Zusammenstellung möglicher Anwendergruppen für TEP analysiert werden:

- Anwender/Käufer
- vertraut/nicht vertraut mit Terminologiearbeit
- erfahren/unerfahren im Umgang mit TEP
- selbstständig/selbstständig in Team/Mitarbeiter kleine/große Firma/Übersetzungsbüro/Organisation/Verwaltungsorgan
- Terminologe/Translator (Übersetzer/Dolmetscher)/Projektmanager/ Lexikograph/Sonstiges
- Hauptsächlich systematische/punktuelle, mono-/bi-/multilinguale, übersetzungsvorbereitende/-nachbereitende/terminologieorientierte, begriffs-/benennungsorientierte Terminologiearbeit

Aus dem t-survey von Zielinski & Ramírez-Safar können folgende Charakteristika für die Anwendercharakterisierung entnommen werden (vgl. 2005:8-25):

- Qualifikation (Arbeitserfahrung in Jahren, Ausbildung/Universitätsabschluss, Zertifizierungen/Zusatzqualifikationen)
- Arbeitsumgebung (Betriebssystem, Art der Internetverbindung)
- Übersetzung (Arbeitssprachen, Sprachkombinationen, Fachbereiche, überwiegende Texttypen, Dateiformate)
- Terminologiearbeit (regelmäßig/auf Anfrage, bilingual/multilingual/monolingual)
- Speicherung von Terminologie (Verwendung von TVS/Word bzw. Excel/in Paralleltexten/Datenbanken/auf Karteikarten, verwendetes TVS [verschiedene kommerzielle Produkte], Charakteristika gewünschter Terminologie, Speicherung von Zusatzinformationen, Anteil der Terminologiearbeit an der Arbeitszeit)

- Verwendung von TEP (cTE aus Ausgangstexten/Paralleltexten/*translation memories*, verwendetes TEP (verschiedene kommerzielle Produkte, Zufriedenheit mit TEP, Desiderata an TEP, möglicher Finanzrahmen für den Erwerb von TEP)

Im EAGLES-Bericht findet sich ein Beispiel für eine Anwendercharakterisierung von CAT-Programmen. Die Entwicklung orientiert sich an Dimensionen von Übersetzungstätigkeiten (Anzahl der Übersetzungen, Anzahl der Arbeitsstunden, Texttypen bzw. -themen, Charakteristika der Originaltexte, Sprachen, Überprüfungsmechanismen für Übersetzungsqualität, Charakteristika der Übersetzungsfirmen bzw. –organisationen) und beschreibt jeweils die unterschiedlichen möglichen Ausprägungen der einzelnen Dimensionen (vgl. 1995:156-162). Daraus abgeleitet ergeben sich weitere Charakteristika:

- Umfang der Terminologiearbeit (Anzahl täglich erfasster Termini)
- Charakteristika der Terminologie (-arbeit) (Korpus- bzw. Texttypen/Überprüfungsmechanismen für Terminologiequalität)
- Charakteristika der Firmen bzw. Organisationen

Auch die Analyse unterschiedlicher Formen der Arbeitsteilung könnte interessant sein: Fleischmann betrachtet z.B. die Beteiligung von Terminologieverantwortlichen, Extraktionsverantwortlichen, Fachgebietsverantwortlichen und weiteren Beteiligten (externe Dienstleister, Übersetzungsabteilung, Praktikanten) an verteilten Workflows (vgl. Fleischmann 2006:64ff.). Wanhoff beschreibt einen verteilten Workflow in einem Übersetzungsunternehmen, in dem neben dem Terminologen auch Übersetzer am Extraktionsprozess und der permanenten Aktualisierung der Terminologie beteiligt sind (vgl. Wanhoff 2007:23).

Es ist gelungen, einige mögliche Charakteristika unterschiedlicher Anwendergruppen zusammenzustellen. Die detaillierte Analyse der einzelnen Charakteristika zur Erstellung von Anwenderprofilen soll jedoch nicht in dieser Arbeit erfolgen, da dazu weitere Daten erhoben werden müssten.

5.4.3 Entwicklung von Anforderungen, Merkmalen & Maßen

Im Rahmen von allgemeinen Adäquatheitsevaluationen müssen die Anforderungen von unbekannten potentiellen Anwendern und Käufern analysiert werden. Dies sollte anhand von Anwenderprofilen erfolgen, die im Rahmen dieser Arbeit leider nicht definiert werden können (vgl. Kap. 5.4.2). Als Inspiration zur Definition von impliziten Anforderungen, die die Anwender selbst nicht formulieren können, sollen folgende Vorschläge empfohlen werden, die überwiegend auf den Beschreibungen des Aufbaus und der Funktionen von TEP (siehe Kap. 2) sowie der praktische Anwendung von TEP (siehe Kap. 3), insbesondere der Probleme im Umgang mit TEP (siehe Kap. 3.3), basieren. Außerdem wurden die Antwortmöglichkeiten des t-survey zur Frage, welche Funktionen die potentiellen Anwender sich von TEP wün-

schen, für die Entwicklung der Ideen zu Rate gezogen (vgl. Zielinski & Ramírez-Safar 2005:23).

Kategorien für die Definition von Anforderungen könnten sein:

- Technische Faktoren (Hardwarevoraussetzungen, Betriebssystem, Kompatibilität, Schnittstellen zu Datenbanken/weiteren Programmen, Installationsaufwand, notwendige Anpassung nach Installation, Eigenständigkeit)
- Vielseitigkeit (verfügbare Sprachen, Möglichkeiten/Grenzen der Anpassung für weitere Sprachen, monolinguale/bilinguale Extraktion)
- Vorverarbeitung (interne/externe Vorverarbeitung, Methoden, notwendige linguistische Ressourcen, mitgelieferte linguistische Ressourcen, Anpassungsmöglichkeiten der Ressourcen, Einstellmöglichkeiten)
- Terminologieidentifikation (Methoden, Qualität des Identifikationsergebnisses, notwendige linguistische Ressourcen, Eingriffsmöglichkeiten im Identifikationsprozess, Verwendung [verschiedener] Stoppwortlisten, automatischer Abgleich mit Terminologiedatenbanken bzw. Ausschlussdatenbanken)
- Vorbereitung (Importdateiformate, Kompatibilität mit Datenbanken, Einstellmöglichkeiten)
- Interne Nachbearbeitung (Art der Repräsentation, Zusatzinformationen, direkte Verlinkung zu Kontextbeispielen, Anzeige der Wortarten, Filter-, Ranking-, Sortier- und Suchmöglichkeiten, Klassifizierungsmöglichkeiten, Auswahl an Informationen zum Export, Möglichkeit der internen Validierung)
- Externe Nachbearbeitung (Schnittstellen zum Export, Exportdateiformate, Art der exportierbaren Zusatzinformationen, Möglichkeiten der Synchronisation mit Datenbanken)
- Bilinguale Terminologieidentifikation (Methoden, Arbeit mit alignierten Texten, sprachübergreifende Identifikation, integrierte Alignierfunktion, Einstellmöglichkeiten, Nachbearbeitungsmöglichkeiten)
- Zusatzfunktionen (Konsistenzprüfung Ausgangstexte/Übersetzung, Terminologieerkennung, mit automatischer Extraktion neuer potentieller Termini, automatische Extraktion von Definitionen, Informationen zu Synonymen, Antonymen, Wortfamilien, etc., Suche nach Kollokationen, Anzeige von semantischen Beziehungen, Erstellung von Ontologien)
- Benutzbarkeit (Manueller Arbeitsaufwand, Komplexität, Erlernbarkeit, Hilfsfunktion, Benutzerhandbuch, Anwendersupport, Peer-to-Peer-Hilfe)

- Arbeitsteilung (Möglichkeiten der Arbeitsteilung, gleichzeitige Bearbeitung, Unterbrechen der einzelnen Prozesse, Bearbeitung einzelner Aufgaben durch unterschiedliche Personen)
- Effizienz & Zuverlässigkeit (Zeitverhalten, Möglichkeiten der gleichzeitigen Nutzung, Stabilität, Stapelverarbeitung)
- Langfristiger Nutzen (Support, bereitgestellte/automatische Updates, Erweiterbarkeit um Zusatzprogramme, voraussichtliche weitere Entwicklung, Marktposition der Firma)
- Kosten (Erwerb, Lizenzen, Support, Schulungen)

Zu Illustrationszwecken wurde eine Analyse der Anforderungen und Merkmale am Beispiel einer spezifischen Evaluation durchgeführt, diese ist in Anhang I dokumentiert. Im nächsten Kapitel erfolgt eine konkretere Entwicklung von Methoden bzw. Orientierungspunkten für die Entwicklung von Methoden.

5.5 Entwicklung von Methoden

Im Rahmen von zahlreichen Publikationen zur Evaluation von TEP wurde der Wunsch nach einer größeren Standardisierung der Evaluationsmethoden geäußert. Zwei Testtypen des EAGLES-Evaluationsrahmen scheinen für die Standardisierung der Evaluation von TEP besonders gut geeignet zu sein: Kriterienüberprüfungen und Benchmarktests. Die EAGLES-Ideen zu diesen beiden Testtypen sollen in diesem Kapitel nocheinmal näher betrachtet werden, um darauf aufbauend neue Ideen und Entwürfe für Methoden bzw. für die Planung von Methoden zur Evaluation von TEP zu entwickeln.

5.5.1 Kriterienüberprüfung

Eine Kriterienüberprüfung ermöglicht eine relativ objektive, standardisierte Evaluation unterschiedlicher Programme desselben Programmtyps (vgl. Kap. 4.4.4.3) und bietet aus diesem Grund eine interessante methodologische Alternative zu gängigen Verfahren für die Evaluation von Terminologieextraktionsprogrammen.

Für die Durchführung einer Kriterienüberprüfung wird eine Kriterienchecklisten erarbeitet, anhand derer die einzelnen TEP auf die enthaltenen Kriterien überprüft werden. Kriterienchecklisten erfordern keine komplizierten Testverfahren, in manchen Fällen reicht die Dokumentation des Programmes zum Ausfüllen der Checkliste (vgl. EAGLES 1995:149). Ein Beispiel für eine Kriterienchecklisten findet sich im EAGLES-Report (1995:171-175), ein weiteres Beispiel für eine solche Checkliste enthält die Arbeit von Saß (2002:120-130, s.a. Kap. 5.3.1).

Eine wichtige Grundlage für die Erstellung einer Kriterienchecklisten ist ein möglichst detaillierter Überblick über die einzelnen Funktionen der zu evaluierenden Programme. Es muss bekannt sein, welche Funktionen die derzeit am Markt erhältlichen Programme aufweisen, in welchen Bereichen Entwicklungen stattfinden und welche Funktionen sich im Umgang als besonders problematisch erweisen (vgl. EAGLES 1995:140). Zunächst müssen alle Programmkomponenten detailliert beschrieben werden und die einzelnen Anwenderanforderungen verdeutlicht werden (vgl. EAGLES 1995:145). Dies ist im Rahmen dieser Arbeit erfolgt: Kapitel 2 und 3 behandeln den Aufbau und Funktionsumfang von TEP sowie die bekannten Anwendungsszenarien, Probleme und Wünsche der Anwender. In Kapitel 5.4.3 wurden anhand der vorhandenen Publikationen Kategorien für mögliche Anforderungen der Anwender ausgearbeitet, diese sind sicherlich noch nicht vollständig, können aber bereits zur Erarbeitung einer neuen – ebenfalls erweiterbaren – Kriterienchecklisten verwendet werden.

Für die Definition der einzelnen Unterpunkte einer Kriterienchecklisten müssen einzelne Merkmale für jede einzelne Funktion definiert werden (vgl. EAGLES 1995:171). Es dürfen nur Merkmale verwendet werden, die ausreichend formalisiert werden können und die zuverlässig (vgl. EAGLES 1995:171) und nicht von Umgebungsfaktoren abhängig sind (vgl. EAGLES 1995:140). Es empfiehlt sich, objektive bewertbare Werte (z.B. binäre Werte wie: vorhanden/nicht vorhanden, bzw. numerische Werte) zu verwenden (vgl. EAGLES 1995: 146). Der Zweck jedes einzelnen Kriteriums sollte erkennbar sein (vgl. EAGLES 1995:140).

Eine Kriterienchecklisten sollte standardisiert sein, das heißt, sie sollte auf jedes TEP angewendet werden können (vgl. EAGLES 1995:140). Nichtsdestotrotz sollte darauf geachtet werden, dass die Kriterienchecklisten es ermöglichen soll, Unterschiede zwischen einzelnen TEP aufzuzeigen (vgl. EAGLES 1995:145). Es sollte Möglichkeiten geben, Innovationen, die bei der Erstellung der Kriterienchecklisten unbekannt werden, in angemessener Form dokumentieren zu können, um besonders innovative Programme nicht zu benachteiligen (vgl. EAGLES 1995:85).

Neue Kriterienchecklisten

Auf der Grundlage der Kriterienchecklisten von Saß (vgl. 2004:120-130) und den vorgeschlagenen Anforderungen aus Kapitel 5.4.3 sowie anhand der im Laufe der Arbeit erarbeiteten detaillierten Funktionen von TEP (siehe Kapitel 2 und 3) wurde eine erweiterte Kriterienchecklisten für eine allgemeine Adäquatheitsevaluationen entwickelt. Diese Checkliste ist als Entwurf zu betrachten ist, der im Rahmen von zukünftigen Evaluationen als Grundlage verwendet werden kann. Der Entwurf wird in Anhang II vollständig wiedergegeben. In diesem Kapitel sollen die einzelnen Bestandteile präsentiert und eine mögliche Anwendung erklärt werden.

Die Checkliste verfügt über 5 Spalten: Im *Kriterienbereich* (Spalte 1) wird das Thema angegeben, zu dem die jeweiligen Kriterien gehören. Als übergeordnete Bereiche wurden die Kategorien *Technische Eignung, Vielseitigkeit, Vorbereitung, Vorverarbeitung, Identifikation, Nachbearbeitung, Zusatzfunktionen, Benutzbarkeit, Arbeitsteilung, Wiederherstellbarkeit, Langfristiger Nutzen* und *Kosten* gewählt. Unterkategorien werden ebenfalls in der ersten Spalte benannt. Die *Position* (Spalte 2) bestimmt die Struktur der Checkliste und identifiziert jedes Kriterium durch eine eindeutige Nummerierung: für jedes untergeordnete Kriterium wird eine weitere Hierarchiestufe verwendet, gleichgeordnete Kriterien finden sich auf der gleichen Stufe. Die *Bedingung* (Spalte 3) wird angegeben, falls das jeweilige Kriterium nur nach der Auswahl eines bestimmten Wertes des übergeordneten Kriteriums verwendet wird. Die Bedingung bezieht sich nicht auf gleichgeordnete Kriterien, sondern jeweils auf das vorhergehende übergeordnete Kriterium. In speziellen Fällen, in denen sich der Wert auf ein anderes vorhergehendes Kriterium bezieht, wird die Position des entsprechenden Kriteriums angegeben. Das Kriterium selbst wird in Form einer Frage unter *Kriterium* (Spalte 4) festgelegt.

Der *Werttyp* (Spalte 5) bezeichnet die den einzelnen Kriterien zugeordnete Werteauswahl. Es können folgende Werttypen zugeordnet werden: Ja/Nein, Nummernfeld (evtl. mit Einheit), Textfeld und Auswahllisten zu den Themen *Benennungsart, Betriebssysteme, bilinguale Identifikation, Browser, Dateiformate Export, Dateiformate Import, Datenbanken, Eigenständigkeit, Extraktionsmethode, Filtereinstellungen, Hilfe, Hinweis Konsistenz, Kompatibilität, Konsistenzprüfung, linguistische Anreicherung, linguistische Einstellungen, linguistische Methoden, linguistische Ressourcen, mono-, bi-, multilingual, Rankingeinstellungen, Schnittstellen, Sprachen, statistische Einstellungen, statistische Methoden, Stoppwort-Einstellungen, Terminologieerkennung neu, Terminologieerkennung vorhanden,Textart, Verfahren, Vorverarbeitungsschritte, Wortart, Zeit, Zusatzinformationen, Zusatzinformationen bilingual.* Die einzelnen Auswahllisten bestehen aus zwei bis zu über dreißig Unterpunkten – eine Übersicht findet sich in Anhang III.

Sowohl die Kriterienbereiche als auch die Kriterien, Werttypen und Auswahllisten sind beliebig erweiterbar – in dieser Liste wurde versucht, alle im Rahmen der Arbeit genannten Faktoren und Unterscheidungsmöglichkeiten zu beachten, die sich für diese Art der Evaluation eigenen – die Qualität des Identifikationsergebnisses wurde bewusst nicht in die Kriterienchecklisteaufgenommen. Eine Evaluation der Qualität des Identifikationsergebnisses erfordert eine aufwendige Planung, Vorbereitung und Durchführung – das Ausfüllen der Checkliste soll ohne die Durchführung komplizierter Textverfahren möglich sein.

Vor einer Anwendung der Checkliste könnten mithilfe von Feldfunktionen die jeweiligen Werttypen z.B. durch Drop-Down-Listen und Ankreuzfunktionen in ein Dokument oder

eine Weboberfläche integriert werden. Darüber hinaus könnten die Bedingungen mit den jeweiligen bedingten Kriterien verknüpft werden, so dass die Anzeige der verwendeten Kriterien automatisiert abläuft. Für eine Evaluation anhand der Checkliste müssen die zu evaluierenden TEP verfügbar sein. Es muss die Möglichkeit bestehen, diese zu installieren und zu verwenden – es sollte für alle TEP dasselbe Setup verwendet werden. Außerdem sollte Testmaterial für die auf primitiven Aktionen basierenden Benchmarktests (s.a. Kap. 5.5.2.1) und die Zeitüberprüfungen vorliegen. Dieses Testmaterial sollte für alle zu evaluierenden TEP gleich sein und in der Evaluationsdokumentation beschrieben werden.

5.5.2 Benchmarktests

Benchmarktests sind eine beliebte und vielfach verwendete Methode im Rahmen von TEP-Evaluationen. Sie können z.B. verwendet werden, um die Funktionalität (v.a. die Richtigkeit) und die Effizienz (Zeitverhalten, Verbrauchsverhalten), sowie die Gebrauchsqualität (Produktivität) zu testen.

Bei Benchmarktests sollte bedacht werden, dass – auch wenn sie gerne genutzt werden, um die interne Qualität eines Programmes zu testen – externe Faktoren wie das jeweilige Setup, der Anwender bzw. die Anwender, zu beachtende Workflows und damit verbundene Datenströme nicht unbeachtet bleiben können (vgl. EAGLES 1995:150f.).

Zur Entwicklung von Benchmarks sollten zunächst Anforderungen an den gewünschten Nutzen des Programmes definiert werden, von denen in weiterer Folge zuverlässige, gültige und durchführbare Benchmarks abgeleitet werden können (vgl. EAGLES 1995:150). Benchmarks bestehen aus dem jeweiligen Merkmal und zugeordnetem Maß, und der Methode zur Messung des Maßes einschließlich des zu verwendenden Testmaterials (zumindest Vorgaben), den Erfahrungsansprüchen an den Evaluator, der jeweiligen Vorgangsweise und den Interpretationsvorgaben. Je nach gewünschter Standardisierung ist eine unterschiedlich detaillierte Definition der Benchmarks möglich: innerhalb der Vorgaben können alle Details festgelegt und das Testmaterial bereits mitgeliefert werden. Jedoch ist es auch möglich, die Definition dem Evaluator zu überlassen und nur grundsätzliche Richtlinien vorzugeben. Benchmarktests sollten ohne die Beteiligung von Anwendern durchgeführt werden, da eine möglichst objektive Durchführung erforderlich ist (vgl. EAGLES 1995:153f.).

5.5.2.1 Primitive Aktionen als Benchmarks

Eine Möglichkeit zur Definition von Benchmarks besteht in der Reduktion der Handlungen auf sogenannte primitive Aktionen (*primitive actions*), wie z.B. Mausklick bzw. Tastendruck. Eine Benchmark kann z.B. beschreiben, wie viele primitive Aktionen getätigt werden müssen, um ein Extraktionsprojekt anzulegen. Beispiele für Benchmarks im Bezug auf *trans-*

lation memories finden sich im EAGLES-Report (vgl. 1995:154f.). Benchmarktests anhand von primitiven Aktionen könnten genutzt werden, um die Benutzbarkeit von TEP zu evaluieren, Beispiele hierfür finden sich in der Kriterienchecklisten in Anhang II im Kriterienbereich Benutzbarkeit.

5.5.2.2 Benchmarktests zur Evaluation des Identifikationsergebnisses

Benchmarktests werden im Rahmen von TEP-Evaluationen gerne genutzt, um die Qualität des Identifikationsergebnisses zu evaluieren. In Folge sollen einige Überlegungen zur Verwendung von Benchmarktests zu diesem Zweck getroffen werden.

Benchmarktests sind experimentelle Evaluationen, die sich durch die Verwendung von objektiven, stark kontrollierten Methoden auszeichen. Die Anwendung eines solchen Tests sollte im Vorfeld detailliert geplant werden, und während der Durchführung sowie nach Abschluß ausreichend dokumentiert werden. Die Dokumentation kann die Nachvollziehbarkeit und unter Umständen auch die Repräsentativität einer Evaluation erhöhen und sollte dementsprechend mit großer Sorgfalt erfolgen.

Einige der in Kapitel 5.1.4 und 5.3.2 beschriebenen Methodologien zur Evaluation des Identifikationergebnisses weisen interessante Ansätze auf, diese sollen hier nochmals zusammengefasst werden.

Validierung

Thurmair schlägt vor, die Problematik der Validierung von Identifikationsergebnissen zu lösen, indem all jene potentiellen Termini bestätigt werden, die semantische Informationsträger sind, und lediglich jene verworfen werden, die unter keinen Umständen als potentielle Termini gelten könnten (vgl. 2003:175). Dadurch wird die Repräsentativität des Ergebnisses eventuell vom eingeschränkten Anwendungsrahmen einer spezifischen Evaluation auf die Repräsentativität einer allgemeinen Evaluation ausgeweitet. Jedoch muss die Vorgabe „semantischer Informationsträger“ im Vorfeld einer Evaluation noch genauer definiert werden, und das Testmaterial sowie sonstige Methoden müssen ebenfalls an die erweiterte Repräsentativität angepasst werden, da diese sonst nicht gewährleistet werden kann.

Für die Validierung der Identifikationsergebnisse werden bei Lieske alle Identifikationsergebnisse zusammengefasst, Wiederholungen bereinigt und eine vollständige Liste aller Identifikationsergebnisse entsteht, in der nicht mehr erkennbar ist, welche Ergebnisse von welchem TEP identifiziert worden sind (vgl. 2002:121). Diese Vorgangsweise erhöht die Objektivität und Effizienz der Validierung sowie die Zuverlässigkeit der Methode, indem garantiert wird, dass keine unterschiedlichen Entscheidungen über gleiche potentielle Benennungen getroffen werden.

Fleischmann entwickelt fünf Kategorien der Korrektheit, um eine detailliertere Betrachtung des Korrektheitsgrades bei der Validierung potentieller Termini zu ermöglichen (vgl. 2006:96ff.). Dadurch kann gewährleistet werden, dass die in binären Klassifizierungen schwierige und oftmals unregelmäßige Zuordnung von uneindeutigen Ergebnissen keine verfälschende Auswirkung auf die Ergebnisse hat.

Eine eigens für die Validierung entwickelte Validierungsumgebung ermöglicht in Lieskes Evaluation eine sehr detaillierte und dennoch formalisierte Validierung (vgl. 2002: 118-121), die darüber hinaus die Effizienz des Validierungsverfahrens vergrößert. Zusätzlich sind weitere Rückschlüsse über die besondere Eignung der Programme für besondere Arten von Termini möglich, z.B. Identifikation von Einwort- im Vergleich zu Mehrwortbenennungen (vgl. Lieske 2002:126f.).

Testkollektionen

Für die Erarbeitung von Testkollektionen empfehlen Vivaldi & Rodríguez, ein Team mit der manuellen Extraktion des gesamten Testmaterials zu betrauen. Jeder Beteiligte soll den vollständigen Text bzw. Korpus bearbeiten. In Folge werden all jene potentiellen Termini als gültig angesehen, die von mindestens einem Teammitglied als Termini ausgewiesen wurden (vgl. Vivaldi & Rodríguez 2007:237f.). Dadurch wird die Beeinflussung der Validierung durch subjektive Entscheidungskriterien eingeschränkt. Als Variante könnte das Team aus drei unterschiedlichen Anwendern von Terminologieextraktionsprogrammen bestehen, z.B. Übersetzer, Terminologen, technische Schreiber.

Um einen geringen Umfang der Testkollektion zu verhindern, verwendet Lieske alle validierten Termini aller im Rahmen der Evaluation verwendeten TEP als Benchmark (vgl. 2002:126). Dadurch wird der Umfang der Testkollektion stark erhöht, jedoch besteht die Gefahr, verfälschte Recall- bzw. Silence-Werte zu erhalten. Zur Sicherstellung der Gültigkeit sollten in diesem Falle mehrere Stichproben zur Überprüfung der Korrektheit der Silence-Werte durchgeführt werden.

Vorgaben

Lieske legt im Vorfeld die unterschiedlichen Blickwinkel fest, aus denen die Qualität der Identifikationsergebnisse betrachtet werden soll: für Anwender werden anschließend Beispiele für Identifikationsergebnisse in den Bericht aufgenommen, für potentielle Käufer statistische Kennzahlen zu quantitativen Ergebnissen. Die Ergebnisse werden jeweils von Erklärungen begleitet, die dem Informationsbedarf der Adressaten entsprechen (vgl. Lieske 2002:121-128).

5.5.2.3 Planung und Entwicklung von Benchmarktests

Anhand dieser Ideen und der starken Betonung der Formalisierung sowie der Bedeutung der Dokumentation wurden einige Fragen erarbeitet, die bei der Planung eines Benchmarktests für die Evaluation von Identifikationsergebnissen beachtet werden sollten:

Allgemein

- Für welchen Zweck soll die Evaluation durchgeführt werden (terminologisch [systematisch/punktuell], translatorisch [übersetzungsvorbereitend/-nachbereitend], etc.)?
- Wie umfangreich sollen die erwarteten Schlussfolgerungen sein? Welche Repräsentativität der Ergebnisse wird angestrebt? Entspricht das verwendete Testmaterial den Anforderungen für die Repräsentativität? (Umfang, Zusammensetzung, Realitätsnähe)
- Welche Einstellungen werden für die Identifikation verwendet? (Standardeinstellungen, in Testläufen als besonders geeignet identifizierte Einstellungen)
- Wird das Testmaterial in einem Zug verarbeitet oder wird es unterteilt? Gibt es besondere Gründe für die Teilung? (Test von einzelnen Funktionen, thematische Unterteilung)

Berechnung von Noise- bzw. Precision-Werten

- Wie erfolgt die Validierung? Wurden ausreichende Vorgaben zur Validierung gemacht? (binär oder klassifizierend, einzeln anhand der Programme oder weiterverarbeitet zu einer anonymisierten vollständigen Liste, durch Terminologen oder fachliche Experten, durch einzelne Validierer oder Teams, in den jeweiligen TEP oder in einer Validierungsumgebung)
- Auf welcher Basis erfolgt die Berechnung? (werden weder vollständig korrekte noch vollständig inkorrekte Ergebnisse als korrekt oder inkorrekt gewertet? Wird zwischen Programmen mit und ohne Grundformenbildung unterschieden?)

Berechnung von Recall- bzw. Silence-Werten

- Soll eine Testkollektion erstellt werden? Welche Repräsentativität der Ergebnisse wird angestrebt? Entspricht die Testkollektion den Anforderungen für die Repräsentativität? (Umfang, Vorgaben zur Auswahl relevanter Termini, Erstellung durch qualifizierte Personen/mehrere Personen, zuverlässige Methode zum Vergleich der manuellen und automatischen Identifikationsergebnisse)
- Sollen alle validierten Ergebnisse zusammen als Benchmark verwendet werden? Welche Repräsentativität wird angestrebt? (Wird durch Stichproben überprüft, ob tatsächlich alle relevanten potentiellen Termini gefunden wurden oder ob die Silence-Werte zu niedrig sind?)

Berechnung des F-Wertes

- Sollen F-Werte berechnet werden? Werden Precision und Noise bzw. Silence und Recall gleich gewichtet oder soll ein Faktor stärkere Auswirkungen haben?

Vorgaben zur Interpretation der Ergebnisse

- Welche Blickwinkel sollen im Bericht beachtet werden? (z.B. Käufer, Anwender) Welche Informationen sind für die jeweiligen Adressaten relevant? Welcher zusätzliche Informationsbedarf besteht zur Erläuterung der Ergebnisse?
- Welche Recall- bzw. Silence- und Precision- bzw. Noise-Werte werden angestrebt? (Mindestvorgaben, Bewertungsschema)

Diese Fragen sollen als Orientierunghilfe bei der Entwicklung eines Benchmarktests für die Analyse des Identifikationsergebnisses dienen. Es sollte dabei eine umfangreiche Planung und Dokumentation (vgl. Kap. 5.2) sowie das Ziel der größtmöglichen Repräsentativität bei der Entscheidungsfindung angestrebt werden, um die Vergleichbarkeit unterschiedlicher Evaluationen zu vergrößern.

5.6 Zusammenfassung

Im Verlauf dieses Kapitels wurden einige Evaluationen von TEP vorgestellt und ihre Methodologie sowie das im Rahmen der Evaluationen verwendete Testmaterial diskutiert. Anschließend wurde die quantitative Analyse der Qualität des Identifikationsergebnisses als häufiger Schwerpunkt dieser Evaluationen betrachtet und einige methodologisch problematischen Aspekte wurden angesprochen. Die nähere Betrachtung zweier auf dem EAGLES-Evaluationsrahmen basierender Evaluationsmethodologien zeigte Perspektiven für die weitere Entwicklung von Evaluationsmethodologien auf. Anhand der in den vorhergehenden Kapiteln erarbeiteten Grundlagen wurden einige Ideen zur Analyse der Evaluationssituation, zu Anwenderprofilen und Anforderungen für die mögliche Entwicklung eines parametrisierbaren Evaluationsmodells gesammelt. Abschließend wurde eine umfassende Kriterienchecklisten für eine allgemeine Kriterienüberprüfung erstellt und innovative Ansätze für die Methodologie von Benchmarktests wurden zu Vorgaben bei der Planung von Benchmarktests zusammengefasst.

6 Zusammenfassung

Publikationen zur computergestützten Terminologieextraktion und zur Evaluation von Terminologieextraktionsprogrammen entstehen oft innerhalb bestimmter Forschungs- und Arbeitsbereiche und reflektieren deshalb nur bestimmte Ansätze und Grundlagen ihrer Disziplin, die nur einen eingeschränkten Blickwinkel auf das Feld der computergestützten Terminologieextraktion zulassen. Ziel dieser Arbeit war es durch eine umfassendere Sichtweise verschiedene Grundlagen und Ansätze darzustellen, um einen „Blick über den Tellerrand“ zu ermöglichen, und anhand des breiten Spektrums unterschiedlicher Herangehensweisen Ideen zu einer möglichen Weiterentwicklung der Evaluationsmethodologie zu sammeln.

Im ersten Kapitel erfolgte eine Diskussion der unterschiedlichen Ziele von Anwendern computergestützter Terminologieextraktion, die zu einer umfassenden Definition geführt hat. Auch die in den unterschiedlichen Publikation(-sbereich)en verwendeten Bezeichnungen wurden vorgestellt und diskutiert. Die Übersicht über die drei „Mutterdisziplinen“ Computerlinguistik, Terminologie und Translationswissenschaft trug ebenfalls zu einer Erweiterung des Blickwinkels bei.

Im zweiten Kapitel wurde versucht, trotz der computerlinguistischen Dominanz in den vorhandenen Publikationen, keine Anwendergruppen aus den Augen zu verlieren. Durch die detaillierte Vorstellung der einzelnen Identifikationsmethoden, der Komponenten und des Aufbaus von Terminologieextraktionsprogrammen wurde ein umfassender Überblick über die einzelnen Methoden und Funktionen von Terminologieextraktionsprogrammen geboten, der die große Diversität in diesem Bereich verdeutlicht.

Für eine weitere Vertiefung wurde im dritten Kapitel die praktische Anwendung von Terminologieextraktionsprogrammen betrachtet. Die einzelnen Phasen und Arbeitsschritte der Anwender zeigten die verschiedenen Interaktions- und Einstellmöglichkeiten, die die unterschiedlichen TEP im Extraktionsprozess bieten, und ermöglichten einen Einblick in die mögliche Beeinflussung der Identifikations- und Extraktionsergebnisse durch die verschiedenen Anwender. Ein besonderer Schwerpunkt lag auf der Diskussion bekannter Probleme, die sich auf zahlreiche Aspekte bezogen und nicht nur die Qualität des Identifikationsergebnisses betrafen.

Im vierten Kapitel wurden die Grundlagen der Evaluation von sprachverarbeitender Software erarbeitet. Als Einstieg wurden Normen zur Softwareevaluation und zur Softwarequalität vorgestellt, in weiterer Folge wurde die Entwicklung von Evaluationsmethodologien anhand des EAGLES-Evaluationsrahmens für die anwenderorientierte Evaluation von sprachverarbeitender Software vorgestellt. Die einzelnen Schritte bis hin zur Entwicklung

von Merkmalen, Maßen und Methoden wurden anhand von Beispielen aus der Terminologieextraktion verdeutlicht. Wichtige Konzepte der Evaluation wie die Gültigkeit und die Zuverlässigkeit von Evaluationsmethodologien wurden thematisiert.

Das fünfte Kapitel widmete sich der Evaluation von Terminologieextraktionsprogrammen. Ein umfassender Überblick über Evaluationen aus terminologischer, translatorischer und computerlingustischer Sicht ermöglichte eine Überprüfung der Methodologien und verwendeten Testmaterialien auf Repräsentativität, Zuverlässigkeit und Gültigkeit. Die Methodologien zur häufig angewendeten Evaluation der Identifikationsergebnisse wurden anschließend auf Probleme und Lösungsansätze untersucht. Zwei auf dem EAGLES-Evaluationsrahmen basierende Evaluationen wurden besonders detailliert betrachtet, die Bedeutung der unterschiedlichen Blickwinkel wurde in diesem Rahmen erneut deutlich.

Aufbauend auf den im Verlauf der Arbeit betrachteten Themen wurden grundlegende Ideen für die Entwicklung eines parametrisierbaren Evaluationsmodells formuliert und Vorschläge für die Analyse der Ziele und Ausgangssitationen, sowie Kategorien für eine mögliche Anwendercharakterisierung und die Definition von Anforderungen entwickelt. Zusätzlich wurden eine umfassende Kriterienchecklisten und Vorgaben für die Planung von Benchmarktests zur Evaluation des Identifikationsergebnisses entwickelt, die im Rahmen von zukünftigen Evaluationen angewendet und weiterentwickelt werden können.

7 Fazit

Die zahlreichen unterschiedlichen Blickwinkel und Interessen bei der Evaluation von Terminologieextraktionsprogrammen haben bisher die Entwicklung von standardisierten bzw. vergleichbaren Evaluationsmethodologien verhindert. Nichtsdestotrotz wünschen zahlreiche Evaluatoren eine stärkere Standardisierung sowie Möglichkeiten zur gemeinsamen Nutzung von Evaluationsmethodologien bzw. einzelnen Methoden und Testmaterialien, um den Aufwand zur Vorbereitung von Evaluationen zu reduzieren und eine größere Vergleichbarkeit der Evaluationsergebnisse zu ermöglichen.

Für eine Weiterentwicklung sind auch die Entwickler von Terminologieextraktionsprogrammen gefordert, die durch eine erhöhte Kompatibilität, mehr Schnittstellen und Exportmöglichkeiten sowie eine verstärkte Nutzung von offenen Datenstandards die Durchführung von Evaluationen sowie die computergestützte Terminologiearbeit erleichtern könnten.

Ein Bewusstsein über die Bedeutung einer breiten Palette an Evaluationskriterien ist in Vorüberlegungen zu Evaluationen deutlich erkennbar, im Verlauf der Evaluation treten jedoch meist zahlreiche Faktoren in den Hintergrund – es bleibt die Qualität des Identifikationsergebnisses als alleiniges Evaluationskriterium. Dieser Entwicklung kann durch den bewussten Einsatz von Methoden zur Abdeckung zahlreicher Evaluationskriterien gegengesteuert werden, die im Rahmen der Arbeit entwickelte Kriterienchecklisteist hierfür geeignet.

Auch die Methodologien zur Evaluation des Identifikationsergebnisses werden den Ansprüchen der Gültigkeit und Zuverlässigkeit nur in einigen Fällen gerecht – eine bewusstere Planung mit besonderem Augenmerk auf objektiven, formalisierten Methoden anhand der in dieser Arbeit entwickelten Fragestellungen könnte in diesem Bereich zu einer Verbesserung führen.

Es gibt zahlreiche weitere Möglichkeiten zur Weiterentwicklung der Evaluationsmethodologie von TEP, ein parametrisierbares Evaluationsmodell könnte angestrebt werden, das den Ansprüchen der Standardisierung unter Berücksichtigung der unterschiedlichen Ziele und Bedürfnisse gerecht wird – als erster Schritt für diese Entwicklung würde sich die systematische Erfassung von Daten zur Entwicklung von Anwenderprofilen anbieten.

Bibliographie

Ahmad, Khurshid & Margaret Rogers (2001): Corpus Linguistics and Terminology Extraction. In: Sue Ellen Wright & Gerhard Budin (Hg.) *Handbook of Terminology Management: Volume 2: Application-Oriented Terminology Management*. Amsterdam: John Benjamins, S. 725-760.

Arntz, Reiner; Picht, Heribert & Felix Mayer (2002): *Einführung in die Terminologiearbeit*. Hildesheim/Zürich/New York: Olms (Studien zu Sprache und Technik 2).

Bøegh, Jørgen (2008): A New Standard for Quality Requirements. *IEEE Software*. Heft 25:2, S. 57-63.

Bowker, Lynne (2008): *Computer-Aided Translation Technology. A Practical Introduction*. Ottawa: University of Ottawa Press.

Cabré, María Teresa (1999): *Terminology: Theory, methods and applications.* Amsterdam/Philadelphia: John Benjamins.

Cabré, María Teresa; Estopá Bagot, Rosa & Jordi Vivaldi Palatresi (2001): Automatic term detection: A review of current systems. In: Didier Bourigault, Christian Jacquemin & Marie-Claude L'Homme (Hg.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins, S. 53-88.

Cabré, María Teresa (2009): Cómo las nuevas tecnologías han modificado la metodología terminológica. In: Carmen Navarro *La Comunicación Especializada: [Congreso Internacional "Terminología, Traducción y Comunicación Especializada", celebrado en Verona en octubre de 2007],* Bern: Peter Lang, S. 147-180.

Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf & Hagen Langer (2004): *Computerlinguistik und Sprachtechnologie.* Heidelberg/Berlin: Spektrum Akademischer Verlag.

Dizdar, Dilek (1999): Skopostheorie. In: Mary Snell Hornby, Hans G. Hönig, Paul Kußmaul & Peter A. Schmitt (Hg.) *Handbuch Translation*. Tübingen: Stauffenburg, S. 104-107.

Duden (2003): *Duden: Deutsches Universalwörterbuch*, 5. Auflage. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

EAGLES (1995): *Evaluation of Natural Language Processing Systems.* EAGLES DOCUMENT EAG-EWG-PR.2 [01.02.10, http://www.issco.unige.ch/pub/ewg95.ps]

EAGLES (1999): *The EAGLES 7-step recipe.* Evaluation Working Group [04.11.09, http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html]

Estopá Bagot, Rosa (1999): *Extracció de terminologia: Elements per la construcció d'un SEACUSE (Sistema d'extracció automàtica de candidats a unitats de significació especialitzada)* (Doktorarbeit). Barcelona: UPF. [07.11.09,

http://www.tesisenxarxa.net/TESIS_UPF/AVAILABLE/TDX-0319102-135659//treb1de2.pdf]

Estrella, Paula; Popescu-Belis Andrei & Margaret King (2009): The FEMTI guidelines for contextual MT evaluation: principles and tools. *Linguistica Antverpiensia New Series*, Heft 8/09. [14.12.09, http://andreipb.free.fr/textes/LANS_8_2009_Estrella_FINAL.pdf]

Evert, Stefan & Arne Fitschen (2004): Textkorpora. In: Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer (Hg.) *Computerlinguistik und Sprachtechnologie.* Heidelberg/Berlin: Spektrum Akademischer Verlag, S. 406-413.

Fleischmann, Klaus (2006): *Termextraktion: Ein Beitrag zum Aufbau einer unternehmensweiten Terminologie* (Masterthese). Krems: Donau-Universität Krems.

Gasser, Yvonne (2004): *Evaluation von Terminologieextraktionsverfahren im Rahmen einer Kosten-/Nutzen-Analyse zur Einführung einer standardisierten Firmenterminologie bei AVL List* (Diplomarbeit). Graz: Institut für Theoretische und Angewandte Translationswissenschaft (ITAT) der Karl-Franzens-Universität Graz.

GIS-Glossar (2006): Glossar des geographischen Informationssystems der Universität Bonn. [07.11.09, http://www.giub.uni-bonn.de/gistutor/anhang/glossar/t.htm]

Gruber, Harald; Körner, Christian; Plösch Reinhold & Stefan Schiffer (2007): Tool Support für ISO 14598 based code quality assessments. In: *Sixth International Conference on the Quality of Information and Communications Technology, QUATIC 2007*. Lissabon, 13.-15. September 2007. Lissabon: IEEE Computer Society, 21-29. [04.04.10, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4335231]

Halama, André (2004): Flache Satzverarbeitung. In: Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer (Hg.) *Computerlinguistik und Sprachtechnologie.* Heidelberg/Berlin: Spektrum Akademischer Verlag, S. 218-231.

Heid, Ulrich (1998): A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology* 5:2, S.161-181.

Heid, Ulrich (2001): Verfahren zur Extraktion von Termkandidaten: ein Überblick. In: Felix Mayer (Hg.) *Dolmetschen und Übersetzen: Der Beruf im Europa des 21. Jahrhunderts*. Freiburg: freigang, mauro + reinke, S. 186-205.

Heid, Ulrich (2004): Computergestützte Lexikographie und Terminologie. In: Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer (Hg.) *Computerlinguistik und Sprachtechnologie.* Heidelberg/Berlin: Spektrum Akademischer Verlag, S. 471-478.

Heid, Ulrich (2006): Extracting term candidates from recursively chunked text. In: Pius ten Hacken (Hg.) *Terminology, Computing and Translation*. Tübingen: Narr, S. 97-115.

Hoffmann, Dirk W. (2008): *Software-Qualität.* Berlin: Springer.

Holz, Florian & Klara Kranebitter (2010): Aufbau und Erweiterung terminologischer Ressourcen unter Verwendung automatischer Terminologieextraktion. *eDITion - das Fachmagazin für Terminologie* 1/2010, S. 20-23. [05.04.10, http://www.iim.fh-koeln.de/dtt/Dokumente/edition_2010_1_web_artikel/edition-2010-1-holz-florian-kranebitter-klara-web.pdf]

Jekat, Susanne & Tanja Schultz (2004): Evaluation sprachverarbeitender Systeme. In: Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer (Hg.) *Computerlinguistik und Sprachtechnologie.* Heidelberg/Berlin: Spektrum Akademischer Verlag, S. 573-590.

Kageura, Kyo; Yoshioka, Masaharu; Takeuchi, Koichi; Koyama, Teruo; Tsuji, Keita & Fuyuki Yoshikane (2001): Recent advances in automatic term recognition: Experiences from the NTCIR workshop on information retrieval and term recognition. *Terminology* 6:2, S. 151-173.

King, Margeret (2005): Accuracy and Suitability: New Challenges for Evaluation. *Language resources and evaluation* 39:1, S. 45-64.

König, Michaela (2009): *Corporate language bei Bene. Übersetzungs- und Terminologiemanagement als Qualitätsfaktoren für die Unternehmenssprache* (Masterarbeit). Wien: Universität Wien.

Langer, Hagen (2004): Syntax und Parsing. In: Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer (Hg.) *Computerlinguistik und Sprachtechnologie.* Heidelberg/Berlin: Spektrum Akademischer Verlag, S. 231-275.

L'Homme, Marie-Claude; Benali, Loubna; Bertrand, Claudine & Patricia Lauduique (1996): Definition of an evaluation grid for term-extraction software. *Terminology* 3:2, S. 291-312.

Lieske, Christian (2002): Pragmatische Evaluierung von Werkzeugen für die Term-Extraktion. In: *eTerminology – Professionelle Terminologie im Zeitalter des Internet. Akten des Symposiums*. Köln, 12.-13. April 2002, Köln: Deutscher Terminologie-Tag e.V., S.109-131.

Mügge, Uwe (2007): Automatische Terminologieextraktion. In: Peter A. Schmitt & Heike E. Jüngst (Hg.) *Translationsqualität*. Frankfurt: Peter Lang, S. 712-726.

Nakagawa, Hiroshi (2001): Experimental evaluation of ranking and selection methods in term extraction. In: Didier Bourigault, Christian Jacquemin & Marie-Claude L'Homme

(Hg.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins, S. 303–325.

Neukirchen, Helmut (2007): Ausgewählte Aspekte der Softwaretechnik: 2. Produktqualität: Qualitätsmodelle (Präsentation). [04.05.10, http://www.swe.informatik.uni-goettingen.de/notes/SS2007/neukirchen/aaswt_02produktqualitaet.pdf]

Pfitzinger, Elmar (2002): *Der Weg von DIN EN ISO 9000 ff. zu total quality management (TQM).* Deutsches Institut für Normung e.V. Berlin/Wien/Zürich: Beuth.

Plösch, Reinhold (2009): Die Bedeutung von Qualität in Softwareprodukten (Präsentation) Softnetkonferenz Hagenberg. [04.04.10, http://www.scch.at/Files/01_Reinhold_Pl%C3%B6sch_Bedeutung%20von%20Qualit%C3%A4t%20in%20Softwareprodukten.pdf]

Reuther, Ursula & Christoph Rösener (2007): CLAT (Controlled Language Authoring Tool) – Qualität von Anfang an! *MDÜ (Mitteilungen für Dolmetscher und Übersetzer)* Heft 2/07, S. 43-44.

Risku, Hanna (1999): Translatorisches Handeln. In: Mary Snell Hornby, Hans G. Hönig, Paul Kußmaul & Peter A. Schmitt (Hg.) *Handbuch Translation*. Tübingen: Stauffenburg, S. 107-112.

Saß, Ricarda (2004): *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).

Sparck-Jones, Karen & Julia R. Galliers (1995): *Evaluating natural language processing systems: An analysis and review*. Berlin/Heidelberg/New York: Springer.

Thurmair, Gregor (2003): Making Term Extraktion Tools Usable. In: *Proceedings of the Conference of Controlled Language Translation, EAMT-CLAW03.* Dublin, 15.-17. Mai 2003, S. 170-179.

Vivaldi, Jorge & Horacio Rodríguez (2006): Some notes about the evaluation of terms and term extraction systems. In: *IV International Conference on Language Resources and Evaluation,* LREC 2006, S. 12-17.

Vivaldi, Jorge & Horacio Rodríguez (2007): Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13:2, S. 225-248.

Wanhoff, Vanessa-Barbara (2007*): Computergestützte Extraktion von Termini aus technischen Texten in italienischer Sprache* (Diplomarbeit). Heidelberg: Ruprecht-Karls-Universität Heidelberg.

Witschel, Friedrich (2005): *Terminologie-Extraktion – Möglichkeiten der Kombination statistischer und musterbasierter Verfahren*. Würzburg: Ergon Verlag.

Wohlfahrt, Nicole (2008): alfaloc - Terminologiemanagement mit System. *eDITion - das Fachmagazin für Terminologie* 2/2008, S. 5-7. [19.10.09, http://www.dttev.org/Dokumente/edition_2008_2_webartikels/01-edition-2008-2-wohlfahrt-web.pdf]

Wright, Sue Ellen & Gerhard Budin (2001): *Handbook of Terminology Management: Volume 2: Application-Oriented Terminology Management*. Amsterdam: John Benjamins.

Zerfaß, Angelika (2006): Terminologieextraktion. *eDITion - das Fachmagazin für Terminologie* 2/2006, S. 21-25. [19.10.09, http://www.iim.fh-koeln.de/dtt/Dokumente/edition_2006_2_web_artikel/edition_2006_2_zerfass.pdf]

Zerfaß, Angelika (2008): Gut geplant ist viel gewonnen – Zyklus der Terminologiearbeit. *MDÜ (Mitteilungen für Dolmetscher und Übersetzer)* Heft 1/08, S. 10-14.

Zielinski, Daniel & Yamile Ramírez-Safar (2005): Research meets practice: t-survey 2005 - An online survey on terminology extraction and terminology management. In: *Proceedings ASLIB 27 - Translating and the Computer*. London, 24.-25. November 2005. London: ASLIB – The Association for Information Management. Heft 27. [19.10.09, http://fr46.uni-saarland.de/download/publs/sdv/t-survey_aslib2005_zielinski.pdf]

Zielinski, Daniel & Yamile Ramírez-Safar (2006): Onlineumfrage zu Terminologieextraktions- und Terminologieverwaltungstools: Wunsch und Wirklichkeit noch weit auseinander. *MDÜ (Mitteilungen für Dolmetscher und Übersetzer)* Heft 3/06, S. 26-29.

Anhang

I. Anforderungen & Merkmale

Dies ist ein Beispiel für die Entwicklung von Anforderungen und Merkmalen für eine spezifische anwenderorientierte Adäquatheitsevaluation. Zunächst soll ein Prozessmodell definiert werden:

> Ein **Assistent** identifiziert mithilfe eines TEP **potentielle Termini** in einem **Textkorpus** und sortiert sie für verschiedene **Experten**, die die **sortierten Termini** im TEP validieren. Der Assistent exportiert die **validierten Termini** in eine **Datenbank**.

Das Prozessmodell wird um Umgebungsfaktoren und Setupinformationen erweitert.

> Ein **Assistent**, der an einem PC mit einem **Windows-Betriebssystem** arbeitet, stellt einen **Textkorpus** aus verschiedenen **HTML-** und **Microsoft Word-Dokumenten** zusammen, identifiziert mithilfe eines TEP **potentielle Termini** für die firmeninterne Terminologie und sortiert sie anhand von **Zusatzinformationen** (in welchem Dokument sie am häufigsten vorkommen) nach Fachgebieten (Wirtschaft, Technik). Verschiedenen **Fachleuten** werden die ihnen zugeordneten **sortierten Termini** angezeigt, sie validieren die **sortierten Termini** anhand weiterer **Zusatzinformationen** (Kontext) und der Assistent exportiert die **validierten Termini** mit **Zusatzinformationen** (Kontext) in eine **Multiterm-Datenbank**. Dort werden die **exportierten Termini** vom **externen Übersetzer** aufgerufen und **übersetzt**. Der Übersetzer betrachtet dazu die **exportierten Kontexte** in der **Datenbank**, er hat keinen Zugriff auf die **Originaldokumente**.

Es werden die folgenden Anforderungen formuliert:

Funktionalität

Das TEP sollte...

... Möglichkeiten bieten, die Vorgaben für relevante Termini zu spezifizieren.

... HTML- und Microsoft Word-Dokumente einlesen können.

... die interne Nachbearbeitung einschließlich einer manuellen Sortierung ermöglichen.

... die interne Validierung ermöglichen.

... Zusatzinformationen (Häufigkeit, Fundstelle und Kontext) anzeigen können.

... über eine Schnittstelle zum Export in eine Multiterm-Datenbank verfügen.

... Zusatzinformationen (Kontext) exportieren können.

... verteilte Workflows unterstützen.

Zuverlässigkeit

Das TEP sollte...

... die gleichzeitige Aktivität mehrere Anwender ermöglichen.

... die Unterbrechung und Wiederaufnahme des Validierungsprozesses unterstützen.

Benutzbarkeit

Das TEP sollte...

... auch für Anwender mit geringen Anwenderkenntnissen einfach bedienbar sein.

... über Hilfsfunktionen verfügen.

Installierbarkeit

Das TEP sollte...

... auf einem PC mit Windows-Betriebssystem installiert werden können.

Aus diesen Anforderungen können folgende Merkmale abgeleitet werden:

1	**Es existieren Möglichkeiten zur Spezifikation der gewünschten Termini in Form von folgenden Parametern:**	
1.1	Mindestfrequenz	Ja/Nein
1.2	Mindestlänge des Token	Ja/Nein
1.3	Maximale Anzahl an Token	Ja/Nein
1.4	Auswahl der zu extrahierenden Wortarten	Ja/Nein
1.5	Auswahl der zu extrahierenden POS-Muster	Ja/Nein
1.6	Möglichkeiten zu Modifizierung der POS-Muster	Ja/Nein
1.7	Sonstiges?	Text
2	**Es können folgende Dateiformate importiert werden:**	
2.1	Microsoft Word-Dokumente, einschließlich *.docx	Ja/Nein
2.2	HTML-Dokumente, einschließlich Filterung der HTML-Tags	Ja/Nein
3	**Es gibt Möglichkeiten der internen Nachbereitung:**	
3.1	Es gibt die Möglichkeit einer internen manuellen Sortierung der Termini	Ja/Nein
3.2	Es gibt die Möglichkeit einer internen Validierung	Ja/Nein
3.2.1	Die Validierung kann unterbrochen und abgespeichert werden	Ja/Nein
3.3	Es können Zusatzinformationen zur Häufigkeit angezeigt werden	Ja/Nein
3.3.1	In welcher Form?	Text
3.4	Es können Zusatzinformationen zur Fundstelle angezeigt werden	Ja/Nein
3.4.1	In welcher Form?	Text
3.5	Es können Kontextbeispiele angezeigt werden	Ja/Nein
3.5.1	In welcher Form?	Text
4	**Es gibt Exportmöglichkeiten:**	
4.1	Es gibt eine Schnittstelle zum Export in eine Multiterm-Datenbank	Ja/Nein
4.1.1	Wenn nicht, welche Exportmöglichkeiten gibt es?	Text
4.2	Es können Kontextbeispiele exportiert werden	Ja/Nein
4.2.1	Wie viele Kontexte können exportiert werden?	Anzahl
4.2.2	Können die zu exportierenden Kontexte ausgewählt werden?	Ja/Nein

5	**Es gibt Hilfsfunktionen:**	
5.1	Es gibt eine programminterne Hilfsfunktion	Ja/Nein
5.2	Es gibt eine Online-Hilfe	Ja/Nein
5.3	Es gibt Anwendersupport	Ja/Nein
5.4	Es gibt Peer-to-Peer-Hilfsgemeinschaften	Ja/Nein
6	**Die Bedienung ist nicht komplex:**	
6.1	Wie viele primitive Aktionen müssen für den gesamten Extraktionsvorgang aus einem Microsoft Word Dokument unter Verwendung der Standardeinstellungen durchgeführt werden?	
	Interpretationsschema	<10 einfach
		10-20 normal
		20-30 komplex
		>30 sehr komplex
6.2	Wie viele primitive Aktionen müssen für die Änderung der Mindestfrequenz mindestens durchgeführt werden?	
	Interpretationsschema	<3 einfach
		3-8 normal
		>8 komplex
6.3	Wie viele primitive Aktionen müssen für die Modifizierung von einem POS-Muster mindestens durchgeführt werden?	
	Interpretationsschema	<5 einfach
		5-10 normal
		>10 komplex
7	**Das TEP unterstützt Arbeitsteilung:**	
7.1	Die einzelnen Phasen (Vorbereitung, Vorverarbeitung, Identifikation, interne Nachbearbeitung) können von unterschiedlichen Anwendern ausgeführt werden	Ja/Nein
7.1.1	Die interne Nachbearbeitung kann unterbrochen und zwischengespeichert werden	Ja/Nein
7.1.1.1	Die unterbrochene Phase der internen Nachbearbeitung kann von anderen Anwendern fortgeführt werden	Ja/Nein
7.2	Es gibt die Möglichkeit des externen Zugriffs auf das TEP	Ja/Nein
7.2.1	Der externe Zugriff ist durch folgende Faktoren eingeschränkt	Text
7.3	Auf das System können gleichzeitig mehrere Anwender zugreifen	Ja/Nein
7.3.1	Wie viele Anwender können maximal gleichzeitig auf das System zugreifen?	Anzahl
8	**Das TEP ist mit dem Setup kompatibel:**	
8.1	Das TEP kann auf einem Windows-PC installiert werden	Ja/Nein
8.1.1	Wenn nein, ist es webbasiert verfügbar?	Ja/Nein

II. Kriterienchecklisten

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
Technische Eignung	**1**			
Hardware-voraussetzungen	1.1		Gibt es besondere Hardwareanforderungen?	Ja/Nein
	1.1.1	falls Ja	Mindestanforderungen Prozessor	Nummernfeld (MHZ)
	1.1.2	falls Ja	Mindestanforderungen Arbeitsspeicher	Nummernfeld (MB RAM)
	1.1.3	falls Ja	Mindestanforderungen Festplattenspeicher	Nummernfeld (MB)
	1.1.4	falls Ja	Sonstiges	Textfeld
Webbasiert	1.2		Ist das TEP webbasiert verwendbar?	Ja/Nein
	1.2.1	falls Ja	Welche Browser können verwendet werden?	Auswahl Browser
Kompatible Betriebssyteme	1.3		Ist das TEP unabhängig von Betriebssystemen?	Ja/Nein
	1.3.1	falls Nein	Welche Betriebssysteme sind kompatibel?	Auswahl Betriebs-systeme
Kompatibilität	1.4		Mit welchen CAT-Programmen ist das TEP kompatibel?	Auswahl Kompatibi-lität
Programmier-schnittstellen	1.5		Gibt es Programmierschnittstellen?	Auswahl Schnittstel-len
Installations-aufwand	1.6		Muss ein Server eingerichtet werden?	Ja/Nein
	1.7		Wie lange dauert eine Installation mit Standardeinstel-lungen?	Auswahl Zeit
	1.8		Wie viele primitive Aktionen müssen von Beginn der Installation bis zur Funktionsfähigkeit des TEP mindestens ausgeführt werden?	Nummernfeld
Eigenständigkeit	1.9		Handelt es sich um ein eigenständiges Programm?	Ja/Nein
	1.9.1	falls Nein	Inwiefern ist die Eigenständigkeit eingeschränkt?	Auswahl Eigenstän-digkeit
Vielseitigkeit	**2**			
Mono-/bilinguale Extraktion	2.1		Welche Arten der Extraktion sind verfügbar?	Auswahl mono-/bilingual
Verfügbare Sprachen	2.1.1	falls monolingu-al	Aus welchen Sprachen ist die Extraktion möglich?	Auswahl Sprachen
	2.1.2	falls bilingual	Aus welchen Ausgangssprachen ist die bilinguale Termi-nologieextraktion möglich?	Auswahl Sprachen
	2.1.3	falls bilingual	Aus welchen Zielsprachen ist die bilinguale Terminologie-extraktion möglich?	Auswahl Sprachen
	2.2		Sind alle Sprachversionen inkludiert	Ja/Nein
	2.2.1	falls Nein	Können zusätzliche Sprachversionen erworben werden?	Ja/Nein
Anpassung für weitere Sprachen	2.3		Kann das Programm für weitere Sprachen angepasst werden?	Ja/Nein
	2.3.1	falls Ja	Müssen neue linguistische Ressourcen bereitge-stellt/erstellt werden?	Ja/Nein
	2.4		Können alle verwendeten linguistischen Ressourcen des Programmes bearbeitet werden?	Ja/Nein
	2.4.1	falls Nein	Welche linguistischen Ressourcen können nicht bearbei-tet werden?	Textfeld
	2.5		Muss der Programmcode für eine Anpassung verändert werden?	Ja/Nein
	2.5.1	falls Ja	Handelt es sich um ein Open Source Programm?	Ja/Nein
Vorbereitung	**3**			
Importdatei-formate	3.1		Welche Dateiformate werden zum Import unterstützt?	Auswahl Datei-formate Import
Kompatibilität mit Datenbanken	3.2		Aus welchen Datenbanken können Daten direkt impor-tiert werden?	Auswahl Daten-banken
Einstellmöglich-keiten	3.3		Können mehrere Dokumente gleichzeitig eingelesen werden?	Ja/Nein

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
	3.3.1	falls Ja	Wie viele Dokumente können maximal gleichzeitig verarbeitet werden?	Nummernfeld
Kapazität	3.4		Welche Datenmenge kann maximal in einem Zug verarbeitet werden?	Nummernfeld
Projektunterstützung	3.5		Unterstützt das TEP die Erstellung von Projekten?	Ja/Nein
	3.5.1	falls Nein	Wie viele primitive Aktionen sind vom Start des Programms bis zum Beginn der Vorverarbeitungsphase notwendig?	Nummernfeld
	3.5.2	falls Ja	Können Projektdaten in Teilmengen verarbeitet werden?	Ja/Nein
Vorverarbeitung	**4**			
interne/externe Vorverarbeitung	4.1		Findet eine interne Vorverarbeitung statt?	Ja/Nein
	4.2		Kann der Text extern vorverarbeitet werden?	Ja/Nein
	4.2.1	falls Ja	Muss der Text extern vorverarbeitet werden?	Ja/Nein
	4.2.1.1	falls Ja	Welche Vorverarbeitungsschritte müssen extern erfolgen?	Auswahl Vorverarbeitungsschritte
Segmentierung	4.3		Findet eine interne Segmentierung statt?	Ja/Nein
	4.3.1.	falls Ja	Kann die Segmentierung individuell angepasst werden?	Ja/Nein
	4.3.1.1	falls Ja	Können die Segmentierungsregeln bearbeitet werden?	Ja/Nein
	4.3.1.1.1	falls Ja	Können zur Segmentierung verwendete linguistische Ressourcen intern bearbeitet werden?	Ja/Nein
	4.3.1.1.2	falls Ja	Können zur Segmentierung verwendete linguistische Ressourcen extern bearbeitet werden?	Ja/Nein
	4.3.2.		Welche Segmentierungsparameter können eingestellt werden?	Textfeld
	4.3.3		Wie viele primitive Aktionen müssen für die Segmentierung mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
Linguistische Anreicherung	4.4		Findet eine linguistische Anreicherung statt?	Ja/Nein
	4.4.1	falls Ja	Welche Verfahren werden zur linguistischen Anreicherung verwendet?	Auswahl linguistische Anreicherung
	4.4.1.1	falls POS-Tagging	Wie viele primitive Aktionen müssen für ein POS-Tagging mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	4.4.1.2	falls Parsing	Wie viele primitive Aktionen müssen für ein Parsing mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	4.4.1.3	falls partielles Parsing	Wie viele primitive Aktionen müssen für ein partielles Parsing mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	4.4.1.4	falls sonstiges Verfahren	Wie viele primitive Aktionen müssen für das Anreicherungsverfahren mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	4.4.2	falls Ja	Welche Anreicherungsparameter können eingestellt werden?	Textfeld
	4.4.3	falls Ja	Können die Ergebnisse der Anreicherung überprüft und bearbeitet werden?	Ja/Nein
Grundformenbildung	4.5		Wird einer Methode zur Grundformenbildung angewendet?	Ja/Nein
	4.5.1	falls Ja	Können die Ergebnisse der Grundformenbildung überprüft und bearbeitet werden?	Ja/Nein
	4.5.2	falls Ja	Wie viele primitive Aktionen müssen für die Grundformenbildung mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
Linguistische Ressourcen	4.6		Sind alle linguistischen Ressourcen für die Vorverarbeitung inkludiert?	Ja/Nein
	4.6.1	falls Nein	Welche zugehörigen linguistischen Ressourcen können hinzugekauft werden?	Auswahl linguistische Ressourcen
	4.6.2	falls Nein	Welche linguistischen Ressourcen müssen erstellt/erworben werden?	Auswahl linguistische Ressourcen
Anpassungsmöglichkeiten der Ressourcen	4.7		Welche linguistischen Ressourcen für die Vorverarbeitung können verändert werden?	Auswahl linguistische Ressourcen

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
Identifikation	**5**			
Monolinguale Identifikations-verfahren	5.1	falls 2.1 mono-lingual	Wird ein linguistisches, statistisches oder hybrides Identi-fikationsverfahren verwendet?	Auswahl Verfahren
Linguistische Methoden	5.1.1	falls linguistisch oder hybrid	Welche linguistischen Methoden werden verwendet?	Auswahl linguisti-sche Methoden
	5.1.2	falls linguistisch oder hybrid	Welche Parameter können für das linguistische Identifi-kationsverfahren eingestellt werden?	Auswahl linguisti-sche Einstellungen
Linguistische Ressourcen	5.1.3	falls linguistisch oder hybrid	Sind alle linguistischen Ressourcen für die linguistischen Identifikationsmethoden inkludiert?	Ja/Nein
	5.1.3.1	falls Nein	Welche zugehörigen linguistischen Ressourcen können hinzugekauft werden?	Auswahl linguisti-sche Ressourcen
	5.1.3.2	falls Nein	Welche linguistischen Ressourcen müssen er-stellt/erworben werden?	Auswahl linguisti-sche Ressourcen
	5.1.4	falls linguistisch oder hybrid	Welche linguistischen Ressourcen für die Vorverarbei-tung können verändert werden?	Auswahl linguisti-sche Ressourcen
	5.1.5	falls linguistisch oder hybrid	Welche Arten von Benennungen können identifiziert werden?	Auswahl Benen-nungsart
	5.1.6	falls linguistisch oder hybrid	Welche Wortarten können identifiziert werden?	Auswahl Wortart
Statistische Methoden	5.1.7	falls statistisch oder hybrid	Welche statistischen Methoden werden verwendet?	Auswahl statistische Methoden
	5.1.8	falls statistisch oder hybrid	Welche Parameter können für das statistische Identifika-tionsverfahren eingestellt werden?	Auswahl statistische Einstellungen
Stoppwortlisten	5.1.9	falls statistisch oder hybrid	Werden Stoppwortlisten verwendet?	Ja/Nein
	5.1.9.1	falls Ja	Können unterschiedliche Stoppwortlisten ausgewählt werden?	Ja/Nein
	5.1.9.1.1	falls Ja	Können unterschiedliche Zeitpunkte für die Verwendung unterschiedlicher Stoppwortlisten definiert werden?	Ja/Nein
	5.1.9.2	falls Ja	Welche Parameter können für die Verwendung von Stoppwortlisten eingestellt werden?	Auswahl Stoppwort-Einstellungen
Heuristiken	5.2		Werden Heuristiken verwendet?	Ja/Nein
	5.2.1	falls Ja	Welche Art von Heuristiken werden verwendet?	Textfeld
	5.2.2	falls Ja	Können die Heuristiken modifiziert werden?	Ja/Nein
	5.2.2.1	falls Ja	Welche Heuristiken können modifiziert werden?	Textfeld
Bootstrapping	5.3		Werden Bootstrapping-Methoden verwendet?	Ja/Nein
	5.3.1	falls Ja	Kann das Bootstrapping-Identifikationsergebnis vor der Wiederverwendung überprüft und bearbeitet werden?	Ja/Nein
	5.3.2	falls Ja	Wie viele Bootstrapping-Phasen gibt es?	Nummernfeld
Terminologie-erkennung	5.4		Können Terminologie- bzw. Ausschlussdatenbanken während der Identifikationsphase zum Abgleich verwen-det werden?	Ja/Nein
	5.4.1	falls Ja	Welche Datenbanken können zur Terminologieerken-nung verwendet werden?	Auswahl Daten-banken
Bilinguale Terminologie-identifikation	5.5	falls 2.1 bilingu-al	Welche Methode verwendet die bilinguale Extraktion?	Auswahl Extrak-tionsmethode
	5.5.1.	falls aligniert	Bietet das TEP eine Komponente zur Alignierung?	Ja/Nein
	5.5.1.1	falls Ja	Kann die Alignierung überprüft und bearbeitet werden?	Ja/Nein
	5.5.2	falls aligniert	Wird für die bilinguale Extraktion zunächst eine mono-linguale Extraktion ausgeführt?	Ja/Nein
	5.5.2.1	falls Ja	Können die Ergebnisse der monolingualen Extraktion überprüft und validiert werden?	Ja/Nein
	5.5.2.2	falls Ja	Siehe monolinguale Terminologieidentifikation ab 5.1	
	5.5.3		Welche zusätzlichen Einstellungen können für die bi-linguale Identifikation gewählt werden?	Auswahl bilinguale Identifikation
Nachbearbeitung	**6**			
Interne Nach-bearbeitung monolinguale Extraktion	6.1	falls 2.1 mono-lingual oder 5.5.2 Ja	Wird das Identifikationsergebnis zur Nachbearbeitung im TEP angezeigt?	Ja/Nein
	6.1.1	falls Nein	Weiter bei Exportfunktion ab 6.2	

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
Zusatzinformationen	6.1.2	falls Ja	Kann das Identifikationsergebnis mit externen Datenbanken abgeglichen werden?	Ja/Nein
	6.1.3	falls Ja	Welche Zusatzinformationen werden bei der Repräsentation der Ergebnisse angezeigt?	Auswahl Zusatzinformationen
	6.1.4	falls Ja	Welche Zusatzinformationen können bearbeitet werden?	Auswahl Zusatzinformationen
	6.1.5	falls Ja	Gibt es eine Filterfunktion bei der internen Nachbearbeitung?	Ja/Nein
	6.1.5.1	falls Ja	Welche Filtereinstellungen können gewählt werden?	Auswahl Filtereinstellungen
	6.1.6	falls Ja	Gibt es eine Rankingfunktion bei der internen Nachbearbeitung?	Ja/Nein
	6.1.6.1	falls Ja	Welche Rankingeinstellungen können gewählt werden?	Auswahl Rankingeinstellungen
	6.1.7	falls Ja	Gibt es eine manuelle Sortiermöglichkeit bei der internen Nachbearbeitung?	Ja/Nein
	6.1.8	falls Ja	Gibt es eine Suchfunktion bei der internen Nachbearbeitung?	Ja/Nein
	6.1.9	falls Ja	Gibt es eine Klassifizierungsfunktion bei der internen Nachbearbeitung?	Ja/Nein
Validierung	6.1.10	falls Ja	Gibt es eine Validierungsfunktion bei der internen Nachbearbeitung?	Ja/Nein
	6.1.10.1	falls Ja	Ist die Validierung obligatorisch?	Ja/Nein
	6.1.10.2	falls Ja	Kann das Validierungsschema frei gewählt werden?	Ja/Nein
	6.1.10.2.1	falls Nein	Gibt es nur binäre Validierungsmöglichkeiten?	Ja/Nein
	6.1.11	falls Ja	Sind die angezeigten Ergebnisse direkt zu Kontextbeispielen im Korpus verlinkt?	Ja/Nein
	6.1.11.1	falls Ja	Wie viele Kontextbeispiele werden maximal angezeigt?	Nummernfeld
	6.1.12	falls Ja	Können durch eine Kontextsuche Kontextbeispiele gesucht werden?	Ja/Nein
Exportvorbereitung monolinguale Extraktion	6.2	falls 2.1 monolingual oder 5.5.2 Ja	Gibt es eine Exportfunktion?	Ja/Nein
	6.2.1	falls Ja	Welche Dateiformate werden zum Export unterstützt?	Auswahl Dateiformate Export
	6.2.1.1	falls Trennzeichenformat	Können die Trennzeichen für Trennzeichendateien individuell bestimmt werden?	Ja/Nein
	6.2.2	falls Ja	In welche Datenbanken können Daten direkt exportiert werden?	Auswahl Datenbanken
Zusatzinformationen	6.2.3	falls Ja	Welche Zusatzinformationen können für den Export ausgewählt werden?	Auswahl Zusatzinformationen
	6.2.4	falls 6.1.5 und 6.2 ja	Können die Ergebnisse in gefiltertem Zustand exportiert werden?	Ja/Nein
	6.2.5	falls 6.1.6 und 6.2 ja	Können die Ergebnisse in gerankter Reihenfolge exportiert werden?	Ja/Nein
	6.2.6	falls 6.1.7 und 6.2 ja	Können die Ergebnisse in manuell sortierter Ordnung exportiert werden?	Ja/Nein
	6.2.7	falls 6.1.8 und 6.2 ja	Können Suchergebnisse einzeln exportiert werden?	Ja/Nein
	6.2.8	falls 6.1.9 und 6.2 ja	Können die klassifizierten Ergebnisse getrennt exportiert werden?	Ja/Nein
	6.2.9	falls Ja	Welche weiteren Zusatzinformationen können manuell für den Export hinzugefügt werden?	Auswahl Zusatzinformationen
	6.2.10	falls 6.1.10 und 6.2 ja	Können nur die validierten Ergebnisse exportiert werden?	Ja/Nein
	6.2.11	falls Ja	Können Kontextbeispiele für den Export ausgewählt werden?	Ja/Nein
	6.2.11.1	falls Ja	Wie viele Kontextbeispiele können maximal exportiert werden?	Nummernfeld
	6.2.11.2	falls Ja	Können diese Kontextbeispiele vor dem Export bearbeitet werden?	Ja/Nein
Interne Nachbearbeitung bilinguale Extraktion	6.3	falls 2.1 bilingual	Gibt es Funktionen zur bilingualen Extraktion im Rahmen der internen Nachbearbeitung?	Ja/Nein
	6.3.1	falls Nein	Weiter bei Exportfunktion ab 6.4	
	6.3.2	falls Ja	Werden mehrere Übersetzungsvorschläge angezeigt?	Ja/Nein

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
	6.3.2.1	falls Ja	Wie viele Übersetzungsvorschläge werden maximal angezeigt?	Nummernfeld
Zusatzinformationen	6.3.3	falls Ja	Welche Zusatzinformationen werden zu den Übersetzungsvorschlägen angezeigt?	Auswahl Zusatzinformationen bilingual
	6.3.4	falls Ja	Welche Zusatzinformationen können bearbeitet werden?	Auswahl Zusatzinformationen bilingual
	6.3.5	falls Ja	Kann das Identifikationsergebnis mit externen Datenbanken abgeglichen werden?	Ja/Nein
	6.3.6	falls Ja	Gibt es eine Filterfunktion für potentielle Terminologiepaare?	Ja/Nein
	6.3.6.1	falls Ja	Welche Filtereinstellungen können gewählt werden?	Auswahl Filtereinstellungen
	6.3.7	falls Ja	Gibt es eine Rankingfunktion für potentielle Terminologiepaare?	Ja/Nein
	6.3.7.1	falls Ja	Welche Rankingeinstellungen können gewählt werden?	Auswahl Rankingeinstellungen
	6.3.8	falls Ja	Gibt es eine manuelle Sortiermöglichkeit für potentielle Terminologiepaare?	Ja/Nein
	6.3.9	falls Ja	Gibt es eine Suchfunktion für potentielle Terminologiepaare?	Ja/Nein
	6.3.10	falls Ja	Gibt es eine Klassifizierungsfunktion für potentielle Terminologiepaare?	Ja/Nein
Validierung	6.3.11	falls Ja	Gibt es eine Validierungsfunktion für potentielle Terminologiepaare?	Ja/Nein
	6.3.11.1	falls Ja	Ist die Validierung obligatorisch?	Ja/Nein
	6.3.11.2	falls Ja	Kann das Validierungsschema frei gewählt werden?	Ja/Nein
	6.3.11.2.1	falls Nein	Gibt es nur binäre Validierungsmöglichkeiten?	Ja/Nein
	6.3.12	falls Ja	Kann in externen Datenbanken nach potentiellen Übersetzungen gesucht werden?	Ja/Nein
	6.3.13	falls Ja	Sind die angezeigten potentiellen Terminologiepaare direkt zu Kontextbeispielen im Korpus verlinkt?	Ja/Nein
	6.3.13.1	falls Ja	Wie viele Kontextbeispiele werden maximal angezeigt?	Nummernfeld
	6.3.13.2	falls Ja	Können durch eine Kontextsuche Kontextbeispiele gesucht werden?	Ja/Nein
Exportvorbereitung bilinguale Extraktion	6.4	falls 2.1 bilingual	Gibt es eine Exportfunktion für Terminologiepaare?	Ja/Nein
	6.4.1	falls Ja	Welche Dateiformate werden zum Export unterstützt?	Auswahl Dateiformate Export
	6.4.1.1	falls Trennzeichendatei	Können die Trennzeichen für Trennzeichendateien individuell bestimmt werden?	Ja/Nein
	6.4.2	falls Ja	In welche Datenbanken können Daten direkt exportiert werden?	Auswahl Datenbanken
Zusatzinformationen	6.4.3	falls Ja	Welche Zusatzinformationen können für den Export ausgewählt werden?	Auswahl Zusatzinformationen bilingual
	6.4.4	falls 6.3.6 und 6.4 ja	Können die Terminologiepaare in gefiltertem Zustand exportiert werden?	Ja/Nein
	6.4.5	falls 6.3.7 und 6.4 ja	Können die Terminologiepaare in gerankter Reihenfolge exportiert werden?	Ja/Nein
	6.4.6	falls 6.3.8 und 6.4 ja	Können die Terminologiepaare in manuell sortierter Ordnung exportiert werden?	Ja/Nein
	6.4.7	falls 6.3.9 und 6.4 ja	Können Suchergebnisse einzeln exportiert werden?	Ja/Nein
	6.4.8	falls 6.3.10 und 6.4 ja	Können die klassifizierten Terminologiepaare getrennt exportiert werden?	Ja/Nein
	6.4.9	falls 6.3.11 und 6.4 ja	Können nur validierte Terminologiepaare exportiert werden?	Ja/Nein
	6.4.10	falls Ja	Können Kontextbeispiele exportiert werden?	Ja/Nein
	6.4.10.1	falls Ja	Können Kontextbeispiele für den Export ausgewählt werden?	Ja/Nein
	6.4.10.2	falls Ja	Wie viele Kontextbeispiele können maximal exportiert werden?	Nummernfeld

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
	6.4.10.3	falls Ja	Können diese Kontextbeispiele vor dem Export bearbeitet werden?	Ja/Nein
Zusatzfunktionen	**7**			
Konsistenzprüfung	7.1		Unterstützt das TEP Konsistenzprüfungsfunktionen?	Ja/Nein
	7.1.1	falls Ja	In welcher Form ist die Konsistenzprüfung vorhanden?	Auswahl Konsistenzprüfung
	7.1.2	falls Ja	Ist die Konsistenzprüfung mono-, bi- oder multilingual?	Auswahl mono-, bi-, multilingual
	7.1.3	falls Ja	Welche Art von Texten kann überprüft werden?	Auswahl Textart
	7.1.4	falls Ja	Kann die Konsistenzprüfung obligatorisch sein?	Ja/Nein
	7.1.5	falls Ja	Kann die Konsistenzprüfung optional sein?	Ja/Nein
	7.1.6	falls Ja	Wie wird auf mangelnde Konsistenz hingewiesen?	Auswahl Hinweis Konsistenz
Terminologieerkennung	7.2		Unterstützt das TEP Terminologieerkennung und Erfassung neuer potentieller Termini?	Ja/Nein
	7.2.1	falls Ja	Können Terminologie- bzw. Ausschlussdatenbanken während der Identifikationsphase zum Abgleich verwendet werden?	Ja/Nein
	7.2.1.1	falls Ja	Welche Datenbanken können zum Abgleich verwendet werden?	Auswahl Datenbanken
	7.2.2	falls Ja	Wie wird auf bereits vorhandene Termini hingewiesen?	Auswahl Terminologieerkennung vorhanden
	7.2.3	falls Ja	Wie wird auf neue potentielle Termini hingewiesen?	Auswahl Terminologieerkennung neu
automatische Extraktion von Definitionen	7.3		Unterstützt das TEP die automatische Erfassung von Definitionen?	Ja/Nein
	7.3.1	falls Ja	Sind die angezeigten potentiellen Definitionen direkt zu Kontextbeispielen im Korpus verlinkt?	Ja/Nein
	7.3.2	falls Ja	Wie viele potentielle Definitionen werden maximal angezeigt?	Nummernfeld
	7.3.3	falls Ja	Können durch eine Kontextsuche Definitionen gesucht werden?	Ja/Nein
	7.3.4	falls Ja	Können Definitionen für den Export ausgewählt werden?	Ja/Nein
	7.3.5	falls Ja	Wie viele Definitionen können maximal exportiert werden?	Nummernfeld
	7.3.5.1	falls Ja	Können diese Definitionen vor dem Export bearbeitet werden?	Ja/Nein
	7.3.6	falls Ja	Ist die Erstellung von Defintionen vor dem Export obligatorisch?	Ja/Nein
Erstellung von Ontologien	7.4		Unterstützt das TEP die Erstellung von Ontologien?	Ja/Nein
Benutzbarkeit	**8**			
GUI-Sprache	8.1		In welchen Sprachen ist die Bedienungsoberfläche verfügbar?	Auswahl Sprachen
Hilfsfunktion	8.2		Welche Arten von Dokumentation sind verfügbar?	Auswahl Hilfe
	8.2.1	falls Benutzerhandbuch	In welchen Sprachen ist das Benutzerhandbuch verfügbar?	Auswahl Sprachen
	8.2.2	falls Benutzerhandbuch	Wie viele Seiten enthält das Benutzerhandbuch?	Nummernfeld
	8.2.3	falls Dokumentation	In welchen Sprachen ist die Dokumentation verfügbar?	Auswahl Sprachen
	8.2.4	falls Hilfsfunktion	In welchen Sprachen ist die Hilfsfunktion verfügbar?	Auswahl Sprachen
	8.2.5	falls Onlinehilfe	In welchen Sprachen ist die Onlinehilfe verfügbar?	Auswahl Sprachen
	8.2.6	falls Anwendersupport	In welchen Sprachen ist der Anwendersupport verfügbar?	Auswahl Sprachen
	8.2.7	falls Anwendersupport	Ist der Anwendersupport kostenlos?	Ja/Nein
	8.2.8	falls Peer-to-Peer-Hilfe	In welchen Sprachen ist die Peer-to-Peer-Hilfe verfügbar?	Auswahl Sprachen
	8.3		Ist jede Funktion des Programmes in irgendeiner Form dokumentiert?	Ja/Nein

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
	8.4		Können die Standardeinstellungen der einzelnen Komponenten verändert werden?	Ja/Nein
	8.5		Können verschiedene Anwendungsprofile erstellt werden?	Ja/Nein
	8.6		Können Menüs und Symbolleisten personalisiert werden?	Ja/Nein
Manueller Arbeitsaufwand	8.7			
Installation	8.7.1		Wie viele primitive Aktionen müssen von Beginn der Installation bis zur Funktionsfähigkeit des TEP mindestens ausgeführt werden?	Nummernfeld
	8.7.2		Wie lange dauert eine Installation mit Standardeinstellungen?	Auswahl Zeit
Vorbereitung	8.7.3		Wie viele primitive Aktionen sind vom Start des Programms bis zum Beginn der Vorverarbeitungsphase notwendig?	Nummernfeld
	8.7.4	falls 3.5 Projektunterstützung	Wie viele primitive Aktionen sind für die Erstellung eines Projektes mindestens notwendig?	Nummernfeld
	8.7.5		Wie lange dauert die Vorbereitung vom Start des Programms bis zum Beginn der Vorverarbeitungsphase mit Standardeinstellungen?	Auswahl Zeit
Vorverarbeitung	8.7.6	falls 4.3 Segmentierung	Wie viele primitive Aktionen müssen für die Segmentierung mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.7	falls 4.4.1 POS-Tagging	Wie viele primitive Aktionen müssen für ein POS-Tagging mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.8	falls 4.4.1 Parsing	Wie viele primitive Aktionen müssen für ein Parsing mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.9	falls 4.4.1 partielles Parsing	Wie viele primitive Aktionen müssen für ein partielles Parsing mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.10	falls 4.4.1 sonstiges Verfahren	Wie viele primitive Aktionen müssen für das Anreicherungsverfahren mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.11	falls 4.5 Grundformenbildung	Wie viele primitive Aktionen müssen für die Grundformenbildung mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.12		Wie viele primitive Aktionen müssen vom Beginn der Vorverarbeitungsphase bis zum Beginn der Identifikationsphase mindestens durchgeführt werden?	Nummernfeld
	8.7.13		Wie lange dauert die Vorverarbeitung vom Beginn der Vorverarbeitungsphase bis zum Beginn der Identifikationsphase mit Standardeinstellungen ohne die Bearbeitung von Ergebnissen?	Auswahl Zeit
Identifikation	8.7.14	falls 5.1 linguistische Identifikation	Wie viele primitive Aktionen müssen für das linguistische Identifikationsverfahren mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.15	falls 5.1 statistische Identifikation	Wie viele primitive Aktionen müssen für das statistische Identifikationsverfahren mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.16	falls 5.1 hybride Identifikation	Wie viele primitive Aktionen müssen für das hybride Identifikationsverfahren mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.17	falls 5.2 Bootstrapping-Identifikation	Wie viele primitive Aktionen müssen für das Bootstrapping-Identifikationsverfahren mit Standardeinstellungen mindestens durchgeführt werden?	Nummernfeld
	8.7.18	falls 5.4.1 Alignierung	Wie viele primitive Aktionen müssen für eine Alignierung mindestens durchgeführt werden?	Nummernfeld
	8.7.19	falls 5.4.1.1 ja	Wie viele primitive Aktionen müssen für eine Alignierung mit der Bearbeitung von einem Eintrag mindestens durchgeführt werden?	Nummernfeld
	8.7.20		Wie viele primitive Aktionen müssen vom Beginn der Identifikationsphase bis zum Export der Ergebnisse mindestens durchgeführt werden?	Nummernfeld
	8.7.21		Wie lange dauert die Identifikation vom Beginn der Identifikationsphase bis zum Export der Ergebnisse mit Standardeinstellungen ohne die Bearbeitung von Ergebnissen?	Auswahl Zeit

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
Gesamt	8.7.22		Wie viele primitive Aktionen müssen vom Start des Programms bis zum Export des Identifikationsergebnisses mit Standardeinstellungen ohne die Bearbeitung von Ergebnissen mindestens durchgeführt werden?	Nummernfeld
	8.7.23		Wie lange dauert die Extraktion vom Start des Programms bis zum Export des Identifikationsergebnisses mit Standardeinstellungen ohne die Bearbeitung von Ergebnissen?	Auswahl Zeit
Arbeitsteilung	**9**			
	9.1		Unterstützt das TEP Arbeitsteilung?	Ja/Nein
	9.1.1	falls Ja	Können die einzelnen Phasen (Vorbereitung, Vorverarbeitung, Identifikation, interne Nachbearbeitung) von unterschiedlichen Anwendern ausgeführt werden?	Ja/Nein
	9.1.2	falls Ja	Können die einzelnen Phasen unterbrochen und zwischengespeichert werden?	Ja/Nein
	9.1.2.1	falls Ja	Können unterbrochene und zwischengespeicherte Phasen von anderen Anwendern fortgeführt werden?	Ja/Nein
Externer Zugriff	9.2		Gibt es Einschränkungen für externe Zugriffe auf das TEP?	Ja/Nein
	9.2.1	falls Ja	Welche?	Textfeld
Mehrere Anwender	9.3		Kann auf das System gleichzeitig von mehreren Anwendern zugegriffen werden?	Ja/Nein
	9.3.1	falls Ja	Wie viele Anwender können maximal gleichzeitig auf das System zugreifen?	Nummernfeld
Wiederherstellbarkeit	**10**			
	10.1		Gibt es eine Rückgängig-Funktion?	Ja/Nein
	10.2		Gibt es eine Wiederherstellfunktion für Daten?	Ja/Nein
	10.3		Gibt es eine Wiederherstellfunktion für Einstellungen?	Ja/Nein
	10.4		Gibt es eine automatische Speicherungsfunktion?	Ja/Nein
	10.5		Gibt es eine Backup-Funktion?	Ja/Nein
Langfristiger Nutzen	**11**			
Aktualität	11.1		Wann wurde die aktuellste Version veröffentlicht?	Nummernfeld (Datum)
Reife	11.2		Wann wurde die erste Version veröffentlicht?	Nummernfeld (Datum)
	11.3		Wie viele Versionen wurden einschließlich der ersten und der aktuellsten Version veröffentlicht?	Nummernfeld
Updates	11.4		Werden Updates bereitgestellt?	Ja/Nein
	11.4.1	falls Ja	Wie viele Updates wurden in den letzten 12 Monaten veröffentlicht?	Nummernfeld
	11.4.2	falls Ja	Müssen Updates gekauft werden?	Ja/Nein
	11.4.2.1	falls Ja	Wieviel kostet das aktuellste Update?	Nummernfeld (Euro)
Entwicklung	11.5		Ist eine neue Version angekündigt?	Ja/Nein
	11.5.1	falls Ja	Für wann ist die neue Version geplant?	Nummernfeld (Datum)
Marktposition der Firma	11.6		Sind Daten zum Marktanteil bekannt?	Ja/Nein
	11.6.1	falls Ja	Welchen Marktanteil hat das TEP?	Nummernfeld (Prozent)
Kosten	**12**			
Versionen/Pakete	12.1		Wie viele unterschiedliche Versionen bzw. Softwarepakete des TEP gibt es?	Nummernfeld
	12.1.1	n mal (laut Nummernfeld 12.1)	Wie heißt die Komponente bzw. das Softwarepaket?	Textfeld
	12.1.2	n mal (laut Nummernfeld 12.1)	Welche Komponenten umfasst die Version bzw. das Softwarepaket?	Textfeld

Kriterienbereich	Position	Bedingung	Kriterium	Werttyp
	12.1.3	n mal (laut Nummernfeld 12.1)	Wie viel kostet die Version bzw. das Softwarepaket?	Nummernfeld (Euro)
Lizenzen	12.2		Müssen einzelne Lizenzen für die Verwendung erworben werden?	Ja/Nein
	12.2.1	falls Ja	Wieviel kostet eine Lizenz für die interne Verwendung minimal?	Nummernfeld (Euro)
	12.2.2	falls Ja	Wieviel kostet eine Lizenz für die externe Verwendung minimal?	Nummernfeld (Euro)
	12.2.3	falls Ja	Wieviel kostet eine Lizenz für die interne Verwendung maximal?	Nummernfeld (Euro)
	12.2.4	falls Ja	Wieviel kostet eine Lizenz für die externe Verwendung maximal?	Nummernfeld (Euro)
Support	12.3		Gibt es ein Supportangebot?	Ja/Nein
	12.3.1	falls Ja	Wie hoch sind die minimalen einjährigen Supportkosten?	Nummernfeld (Euro)
	12.3.2	falls Ja	Wie hoch sind die maximalen einjährigen Supportkosten?	Nummernfeld (Euro)
Schulungen	12.4		Werden Schulungen angeboten?	Ja/Nein
	12.4.1	falls Ja	Wie hoch sind die minimalen Schulungskosten pro Person?	Nummernfeld (Euro)
	12.4.2	falls Ja	Wie hoch sind die maximalen Schulungskosten pro Person?	Nummernfeld (Euro)

III. Auswahllisten

Benennungsart	Betriebssysteme	bilinguale Identifikation	Browser
Einwortbenennungen	Alle	maximale Anzahl potentielle Übersetzungen	Alle
Mehrwortbenennungen	Windows XP	minimale Häufigkeit Terminologiepaare	Internet Explorer
	Windows 2000	Sonstiges	Mozilla Firefox
	Windows Vista		Opera
	Windows 7		Safari
	Windows Server 2003		Chrome
	Windows Server 2008		Sonstige
	64-bit-Versionen		
	Linux		
	Mac OS		
	Sonstige		

Dateiformate Export	Dateiformate Import	Datenbanken
Extensible Markup Language (XML)	Active Server Pages (ASP)	acrolinx Terminologiemanager
Machine-Readable Terminology Interchange Format (MARTIF)	Active Server Pages (ASPX)	crossTerm
Microsoft Office Excel (XLS)	Adobe InDesign (*.isc)	Déjàvu
Microsoft Office Excel (XLSX)	Ausführbare Dateien (EXE)	DataTerm
Microsoft Office Excel (XLT)	Darwin Information Typing Architecture (DITA)	Docuterm
Microsoft Office Powerpoint (POT)	Extensible Markup Language (XML)	JDBC-kompatible Datenbanken
Microsoft Office Powerpoint (PPS)	HyperText Markup Language (HTML)	Lingo
Microsoft Office Powerpoint (PPT)	IBM Translation Manager (EXP)	Lookup
Microsoft Office Powerpoint (PPTX)	JavaServer Pages (JSP)	MS Access
Microsoft Word (DOC)	Maker Interchange Format (MIF)	ODBC-kompatible Datenbanken
Microsoft Word (DOCX)	Microsoft Office Excel (XLS)	SDL Multiterm
Nur Text (TXT)	Microsoft Office Excel (XLSX)	Termstar
OASIS Open Document Format (ODF)	Microsoft Office Excel (XLT)	Termweb
Open Lexicon Interchange Format (OLIF)	Microsoft Office Powerpoint (POT)	Termxplorer
Portable Document Format (PDF)	Microsoft Office Powerpoint (PPS)	UniLex IDS
PostScript Format (PS)	Microsoft Office Powerpoint (PPT)	UniTerm Professional
TermBase eXchange (TBX)	Microsoft Office Powerpoint (PPTX)	WebTerm
Translation Memory eXchange Format (TMX)	Microsoft Word (DOC)	Wordbank
Trennzeichenformat (CSV)	Microsoft Word (DOCX)	X-Term
Sonstiges	OASIS Open Document Format (ODF)	Sonstige
	Portable Document Format (PDF)	
	PostScript Format (PS)	
	QuarkXPress (*.qsc, *.txg,*.ttg, *.tag)	
	Ressource File Format (RESX)	
	Ressource Script File Format (RC)	
	Rich Text Format (RTF)	
	Scalable Vector Graphics (SVG)	
	SDL Translation Memory Workbench (TMW)	

Dateiformate Export	Dateiformate Import	Datenbanken
	SDLX Translation Document (ITD)	
	SDLX Translation Memory (MDB)	
	Standard Generalized Markup Language (SGML)	
	TRADOStag Dateiformat (TTX)	
	Translation Memory eXchange Format (TMX)	
	Unformatierter Text (TXT)	
	Sonstiges	

Eigenständigkeit	Extraktionsmethode	Filtereinstellungen	Hilfe	Hinweis Konsistenz
Teil eines Softwarepakets	Aligniert	Abgleich interne Datenbank	Benutzerhandbuch	Markierung im Text
Funktion eines Programms	Sprachübergreifend	Abgleich Stoppwortliste	Dokumentation	Warnmeldung
Sonstiges	Sonstige	Mehrwortbenennungen	Hilfsfunktion	Konsistenzreport
		Zeige validierte	Onlinehilfe	Sonstiges
		Zeige nicht validierte	Anwendersupport	
		Sonstiges	Peer-to-Peer-Hilfe	
			Sonstige	

Kompatibilität	Konsistenzprüfung	linguistische Anreicherung	linguistische Einstellungen
Acrolinx	Funktion des TEP	Pos-Tagging	Suchergebnisse nummerieren
Across	Funktion einer Daten-bank	Parsing	maximale Anzahl Token in Mehrwort-benennung
Araya	Eigenständiges Programm	Zwischenstufen wie partielles Parsing	Auswahl vorbestimmter Muster
Déjàvu	Sonstiges	Sonstiges	Freie Musterbestimmung
MemoQ			Unbekannte Wörter wie Nomen behandeln
OmegaT			Alle Token kleinschreiben
SDL Trados			Höhe Qualitätsfilter
StarTransit			Dateipfad einfügen
Wordfast			Sonstiges
Sonstiges			

linguistische Methoden	linguistische Ressourcen	mono-, bi-, multilingual	Rankingeinstellungen	Schnittstellen
intrinsische Analyse von Einwortbenennungen	Regelsätze	monolingual	Häufigkeit	Funktionsorientiert
intrinsische Analyse von Mehrwortbenennungen	Lexika	Bilingual	Alphabet	Dateiorientiert
extrinsische Analyse	Vergleichsdatenbanken	multilingual	Sonstiges	Objektorientiert
Sonstiges	interne Datenbank			Protokollorientiert

Sprachen	statistische Einstellungen	statistische Methoden	Stoppwort-Einstellungen
Alle	Mindestfrequenz	intrinsische Häufigkeit	Verwendung Stoppwortliste
Deutsch	Mindestlänge Token	extrinsische Häufigkeit	Verwendung alle Stoppwortlisten
Englisch	Maximale Länge Token	Mindestfrequenz	Verwendung bestimmte Stoppwortlisten
Französisch	Maximale Anzahl	absolute Frequenz	Zeitpunkt Verwendung einzelner Liste
Spanisch	Höhe Qualitätsfilter	relative Frequenz	Sonstiges
Italienisch	Dateipfad einfügen	Mutual Informarion	
weitere EU-Sprachen	Suchergebnisse nummerieren	Kookkurenz	

Sprachen	statistische Einstellungen	statistische Methoden	Stoppwort-Einstellungen
Japanisch	maximale Anzahl Token in Mehrwortbenennung	Sonstige	
Chinesisch	Sonstiges		
Sonstige			

Terminologieerkennung neu	Terminologieerkennung vorhanden	Textart	Verfahren
Markierung in Ergebnisanzeige	Markierung in Ergebnisanzeige	Originaltext	Linguistisch
Unbestätigtes Hinzufügen	Unbestätigtes Löschen	Übersetzung	Statistisch
Anfrage um Aufnahmebestätigung	Anfrage um Löschbestätigung		Hybrid
Gesonderte Anzeige/Ausgabe	Gesonderte Anzeige/Ausgabe		Sonstiges
nur auf Abfrage	nur auf Abfrage		
Sonstiges	Sonstiges		

Vorverarbeitungsschritte	Wortart	Zeit	Zusatzinformationen	Zusatzinformationen bilingual
Segmentierung	Nomen	bis 10 Minuten	Themenbereich	Häufigkeit Übersetzung
linguistische Anreicherung	Nominalphrasen	bis 20 Minuten	Dateipfad	Häufigkeit Terminologiepaar
Grundformenbildung	Adjektive	bis 30 Minuten	Suchergebnisnummer	Anzahl potentielle Übersetzungen
Sonstige	Verben	bis 1 Stunde	Häufigkeit	Themenbereich
	Adverbien	bis 1,5 Stunden	validiert/nicht validiert	Dateipfad
	Sonstiges	bis 2 Stunden	Synoynme	Suchergebnisnummer
		bis 2,5 Stunden	Akronyme	validiert/nicht validiert
		länger als 3 Stunden	Wortart	Synoynme
			Definition	Akronyme
			Hinweis	Wortart
			Kontext	Definition
			semantische Beziehungen	Hinweis
			Sonstiges	Kontext
				Semantische Beziehungen
				Sonstiges

Abstract (deutsch)

Diese Masterarbeit behandelt methodologische Aspekte der Evaluation von Terminologieextraktionsprogrammen mit einem besonderen Schwerpunkt auf der anwenderorientierten Adäquatheitsevaluation.

Als Grundlagen dienen ein Forschungsüberblick zur computergestützten Terminologieextraktion aus computerlinguistischer, terminologischer und translatorischer Sicht sowie eine detaillierte Betrachtung der Terminologieextraktionsprogramme, insbesondere ihres Aufbaus und der Funktionsweise von Terminologieidentifikationsmethoden. Der computergestützte Extraktionsprozess wird anhand der einzelnen Aufgaben und zugehörigen Funktionen beschrieben. In diesem Rahmen wird gesondert auf Probleme bei der Anwendung von Terminologieextraktionsprogrammen eingegangen – die Qualität von Identifikationsergebnissen erweist sich als dominantes, jedoch nicht als einziges Thema.

Nach einer Einführung in die Konzepte der Softwarequalität und -evaluation wird der EAGLES-Evaluationsrahmen für sprachverarbeitende Programme ausführlich behandelt. Die einzelnen Phasen der Entwicklung einer Evaluationsmethodologie für eine anwenderorientierte Adäquatheitsevaluation und wichtige Charakteristika der Anforderungen, Merkmale und Methoden sowie der Materialien werden vorgestellt.

Mit einem besonderen Schwerpunkt auf der anwenderorientierten Sichtweise werden Methodologien zur Evaluation von Terminologieextraktionsprogrammen betrachtet und auf problematische Methoden und innovative Ansätze untersucht. Die Methodologie zur Evaluation der oft im Vordergrund stehenden Qualität des Identifikationsergebnisses wird besonders detailliert behandelt, darüber hinaus werden zwei auf dem EAGLES-Evaluationsrahmen basierende Methodologien präsentiert und diskutiert.

Abschließend werden einige Ideen zur möglichen Entwicklung eines parametrisierbaren Evaluationsmodells für anwenderorientierte Adäquatheitsevaluationen von Terminologieextraktionsprogrammen ausgearbeitet und eine umfassende Kriterienchecklistе sowie Empfehlungen für die Planung von Benchmarktests zur Evaluation des Identifikationsergebnisses entwickelt.

Abstract (english)

In the present Master's Thesis, methodological aspects of the evaluation of terminology extraction software are discussed, especially focussing on user-oriented adequacy evaluation.

As a basis, an overview of the current state of research in computer-assisted terminology extraction respecting the different points of view of computational linguists, terminologists and translators is given. Subsequently, the composition of terminology extraction software, especially the functioning of different methods for terminology identification, is outlined. Ensuing, the computer-assisted extraction process is ventilated, including a description of individual tasks and corresponding features as well as a discussion of the problems related to the practical application, being the quality of identification results the predominant, though not the only topic.

After an introduction to concepts of software quality and evaluation, the focus is set on the EAGLES evaluation framework for language processing software. The individual steps in creating an evaluation methodology for user-oriented adequacy evaluations and important characteristics of requirements, attributes and methods as well as of materials are presented.

Stressing the user-oriented point of view, evaluation methodologies for terminology extraction software are examined, highlighting problematic methods and innovative approaches. The evaluation methodology of the often predominant quality of identification results, as well as two examples of evaluation methodology which are based on the EAGLES evaluation framework are given special attention.

Finally, some ideas concerning the possible development of a parametrisable evaluation model for user-oriented adequacy evaluations of terminology extraction software are elaborated, and a comprehensive feature checklist as well as some recommendations for designing benchmark tests for the evaluation of identification results are presented.

Lebenslauf — Katrin Mahlstedt, Bakk. phil.

Kontaktdaten & Biographisches

Adresse	Albrechtsbergergasse 27/4, 1120 Wien
Mobiltelefon	+43/650/7250577
Nationalität	Deutsch
geboren am	25.05.1984
geboren in	Wildeshausen, Deutschland
Familienstand	ledig

Ausbildung & Interkulturelles

im September 2009	Modulprüfung Konferenzdolmetschen
im April 2009	Modulprüfung Fachübersetzen
seit Oktober 2007	Masterstudium Übersetzen & Masterstudium Dolmetschen, Deutsch, Spanisch & Tschechisch, Universität Wien
Okt. 04 - März 07	Bakkalaureatsstudium Übersetzen und Dolmetschen Deutsch, Spanisch & Tschechisch, Universität Wien
im Juni 2003	Abitur am Gymnasium Hamburger Straße, Bremen
2001 - 2003	Gymnasium in Bremen, Deutschland
2001 - 2007	Trainings für interkulturelle Kommunikation & Projektmanagement bei AFS Interkulturelle Begegnungen e.V.
Aug. 00 - Juli 01	Schüleraustausch in Neuquén, Argentinien
1990 - 2000	Grundschule & Gymnasium in Wildeshausen, Deutschland

Ehrenamt & Berufliches

seit 2010	Medienauswertung & Übersetzung, Meta Communication International GmbH, Wien
seit 2009	freiberufliche Übersetzerin & Dolmetscherin
2008 - 2010	freiberufliche Trainerin für den Europäischen Freiwilligendienst
April 07 - Juni 07	Volontariat im Österreichischen Kulturforum in Prag
Aug. 03 - Juli 04	Europäischer Freiwilligendienst in Chrudim, Tschechien
2001 - 2009	Ehrenamtliche Tätigkeit für AFS Interkulturelle Begegnungen e.V.