



Das Konzept einer strikt benutzerorientierten Evaluierung maschineller Übersetzungssysteme (am Beispiel des Projekts MIROSLAV)

Dr. Jutta Marx, Dr. Bernhard Staudinger

FG Informationswissenschaft
Universität Regensburg
93040 Regensburg
e-mail: jutta.marx@sprachlit.uni-regensburg.de

Inhalt

- 1 Vorbemerkung
- 2 Das Projekt MIROSLAV
- 3 Der MIROSLAV-Anwendungspartner
- 4 Das MIROSLAV-Evaluierungskonzept
 - 4.1 Benutzerorientierte Evaluierung
 - 4.2 Einzelevaluierung
 - 4.3 Relevanz- und Informativübersetzung
 - 4.4 Übersetzungsqualität und -produktivität
 - 4.5 Globaler Evaluierungsprozeß
 - 4.6 Black-Box-Ansatz + Lexikonfenster
 - 4.7 Häufigkeit der Evaluierung
 - 4.8 Evaluierungsziele: Qualität
 - 4.9 Evaluierungsziele: Produktivität
- 5 Zusammenfassende Übersicht über die MIROSLAV-Testverfahren
- 6 Ausblick
- 7 Literatur

1 Vorbemerkung

Maschinelle Übersetzungssysteme rücken in letzter Zeit wieder vermehrt in das Interesse von Wissenschaft und Industrie. Nachdem sie lange Zeit als unbrauchbar galten, finden sie mehr und mehr Absatz, nicht zuletzt aufgrund der Tatsache, daß Entwickler und Vermarkter abrücken von dem Anspruch, perfekte Reinüber-



setzungen liefern zu können, die mit Humanübersetzungen vergleichbar sind. Stattdessen enthalten maschinelle Übersetzungssysteme mittlerweile ganz selbstverständlich Posteditionshilfen, die eine komfortable Überarbeitung der von der Maschine gelieferten Rohübersetzung erlauben. Für den Benutzer bleibt (wie immer) das Problem, aus der Gesamtpalette das für seine Bedürfnisse passende Produkt auszuwählen. Die Beurteilung der Qualität eines Systems gestaltet sich dabei jedoch als ausgesprochen schwierig, da eine Vielfalt von beeinflussenden Parametern wie Textart, Verwendungszweck, Zeitdruck u.ä. eine allgemeingültige Einschätzung eines Systems unmöglich macht. Evaluationen des Vermarkters sind dabei ebenso eingeschränkt nutzbar wie die der Entwickler, da erstere natürlich bemüht sind, die Vorteile ihres Systems herauszustreichen, letztere hauptsächlich an der Verbesser- und Erweiterbarkeit der Software interessiert sind, die dem Benutzer in seiner konkreten Arbeitssituation wenig weiterhelfen. Evaluationen unabhängiger Dritter wie z.B. Computerzeitschriften mangelt es dagegen an Konkretheit. Da sie nicht auf den speziellen Anwendungskontext eines potentiellen Benutzers eingehen können, sind ihre Bewertungen höchstens als erste Orientierungshilfe zu verstehen.

Das vorliegende Papier versucht diese Lücke zu schließen, indem es ein Konzept einer benutzerorientierten Evaluation für maschinelle Übersetzungssysteme vorstellt, das das zu beurteilende Übersetzungssystem aus der Sicht eines potentiellen Benutzers untersucht. Es wird somit versucht, konkrete Aussagen über die Angemessenheit und den Nutzen eben dieses Systems für den Benutzer in seiner speziellen Arbeitssituation zu machen. Generelle Beurteilungen zum Leistungsspektrum eines Systems bzw. zu seiner Erweiter- oder Verbesserbarkeit, d.h. diagnostische Aussagen interessieren in diesem Zusammenhang weniger.

Die gestellten Aufgaben sind Teil des Projektes MIROSLAV (Machine Translation Initiative for Russian and other Slavic Languages).

2 Das Projekt MIROSLAV

MIROSLAV ist ein von BMBF gefördertes Verbundprojekt zwischen der Gesellschaft für multilinguale Systeme (GMS) in Berlin, Sietec/München, dem Institut für Slavistik der Humboldt-Universität zu Berlin und der FG Informationswissenschaft der Universität Regensburg (IWR). Hauptziel von MIROSLAV ist zunächst die Entwicklung eines maschinellen Übersetzungssystems für das Sprachpaar Russisch - Deutsch; andere slavische Sprachen sollen folgen.

Die Aufgaben der IWR bestehen dabei in der Erarbeitung von lexikalisch-semanticen Lösungen für komplexe Transfers und der Evaluation des russisch-deutschen Übersetzungssystems, das momentan als Unix-basierter Prototyp auf der Basis des Forschungssystems METAL existiert. Für Herbst 1998 ist ein PC-System angekündigt, das im Rahmen der T1-Linie von Langenscheidt vermarktet wird.

Als Anwendungspartner für die Evaluierungsarbeiten konnte die Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS), insbesondere die Abteilung „Informationstransfer Osteuropa“ in Berlin (kurz: GESIS Berlin) gewonnen werden.

3 Der MIROSLAV-Anwendungspartner

Die Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. unterhält seit 1992 eine Außenstelle in Berlin, die sich hauptsächlich mit der Bestandssicherung von Ergebnissen der sozialwissenschaftlichen Forschung in der ehemaligen DDR sowie dem Informationstransfer von und nach Osteuropa zu sozialwissenschaftlichen Themen beschäftigt. Darüber hinaus berät sie in Fragen sozialwissenschaftlicher Methodik sowie zu den Möglichkeiten der Informationsgewinnung in den neuen Bundesländern und in Osteuropa.

Im Rahmen des Informationstransfers zwischen den ost- und westeuropäischen Ländern führt die GESIS Berlin regelmäßige Erhebungen zu Instituten, Zeitschriften und Projekten in Osteuropa durch. Die hierdurch gewonnenen Daten werden durch Informationen aus Printmaterialien wie z.B. Zeitschriften erweitert und publiziert (z.B. im Newsletter „Sozialwissenschaften in Osteuropa“). Ein Teil dieser Informationen geht in die Bestände der vom Informationszentrum Sozialwissenschaften in Bonn (kurz: IZ Bonn) unterhaltenen Datenbank zu sozialwissenschaftlichen Forschungsprojekten (kurz: FORIS) bzw. in die von der GESIS Berlin selbst unterhaltene Datenbank zu sozialwissenschaftlichen Institutionen in Osteuropa ein (kurz: Institutionendatenbank). (Nähere Informationen zur GESIS Berlin s. Marx/Mutschke/Schommler 1995 sowie die WWW-Homepage der GESIS-Berlin unter <http://www.berlin.iz-soz.de/>.)

Zur Wahrung dieser Aufgaben fällt bei der GESIS Berlin zahlreiches Textmaterial in russischer Sprache an, das Gegenstand eines Übersetzungssystems sein könnte. Es handelt sich dabei im wesentlichen um folgende Textsorten¹:

- Im Rahmen einer bisher alle zwei Jahre durchgeführten Umfrage zu laufenden, geplanten oder bereits abgeschlossenen Forschungsprojekten in Osteuropa (d.i. Teil der sog. FORIS-Umfrage) bzgl. sozialwissenschaftlicher Problemstellungen sind von der GESIS Berlin auch zahlreiche Fragebögen in russischer Sprache auszuwerten.
- Zeitschriften, Aufsätze, Monographien und Jahresberichte in russischer Sprache werden regelmäßig von der GESIS Berlin in Hinblick auf Konferenzberichte, Literatur- und Forschungsprojekthinweise, Institutsbeschreibungen u.ä. ausgewertet.
- Ebenfalls für Institutsprofile, die u.a. im Newsletter „Sozialwissenschaften in Osteuropa“ dargestellt werden oder in die Institutionendatenbank eingehen, werden Texte im Internet wie z.B. die Homepages osteuropäischer Institute herangezogen.

¹ Die nicht maschinenlesbaren Texte wie FORIS-Fragebögen oder Zeitschriften müssen zuvor mittels Scanning und OCR erfaßt werden. Hierbei konnten mit dem Softwarepaket FineReader Professional 3.0 von ABBYY Software House bereits sehr gute Ergebnisse erzielt werden.

4 Das MIROSLAV-Evaluierungskonzept

Nach eingehender Sichtung der in der Literatur verfügbaren Evaluierungstypologien und -konzepte erschien das Schema nach Bourbeau 1990 am besten geeignet, die Projektziele von MIROSLAV zu repräsentieren. Das Bourbeau'sche Schema ist hierarchisch gegliedert und somit übersichtlich und erlaubt eine schrittweise Verfeinerung der angestrebten Ziele. Es umfaßt zudem die Hauptkriterien der anderen „klassischen“ Evaluierungsstrategien von Van Slype 1979, Vasconcellos 1988 und Way 1994 bzw. kann problemlos um fehlende Punkte erweitert werden.

Nachstehende Abbildung zeigt das Evaluierungskonzept in MIROSLAV, aufgebaut analog zum hierarchischen Schema nach Bourbeau, das um zwei Grundsatzentscheidungen ergänzt (d.i. benutzerorientierte Evaluierung und Einzelevaluierung) sowie um die für das Projekt wesentlichen Evaluierungsziele konkretisiert wurde.

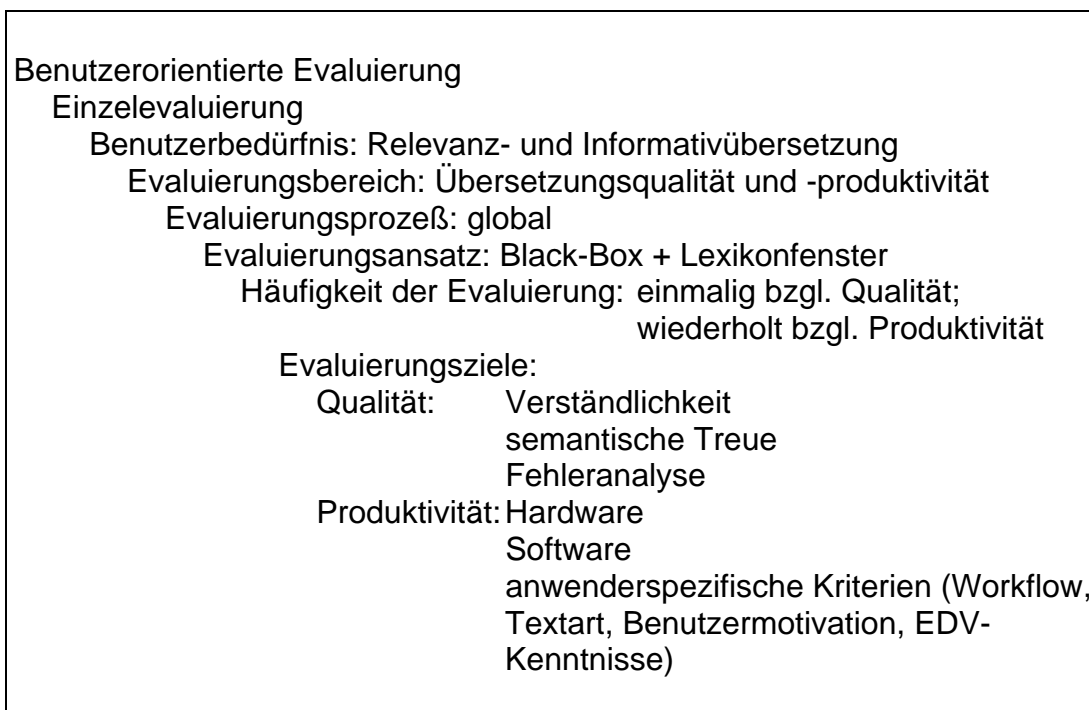


Abb. 1: Das Evaluierungskonzept in MIROSLAV (angelehnt an Bourbeau 1990:43)

Im folgenden werden nun die einzelnen Ebenen dieser Evaluierungshierarchie genauer erläutert sowie die konkreten Evaluierungsziele detailliert beschrieben.

4.1 Benutzerorientierte Evaluierung

In MIROSLAV wird eine streng benutzerorientierte Evaluierung angestrebt, die das zu beurteilende MT-System aus der Sicht eines potentiellen Benutzers untersucht². Es wird somit versucht, konkrete Aussagen über die Angemessenheit und

² Die Abkehr von einer generellen Evaluierung hin zu einer aufgabenorientierten Evaluierung auf dem Gebiet der Sprachverarbeitung und insbesondere der maschinellen Übersetzung setzt sich in der Community immer mehr durch (s. Rubio/Gallardo/Castro/Tejada

den Nutzen eben dieses Systems für den Benutzer in seiner speziellen Arbeitssituation zu machen. Generelle Beurteilungen zum Leistungsspektrum eines Systems bzw. zu seiner Erweiter- oder Verbesserbarkeit, d.h. diagnostische Aussagen interessieren in diesem Zusammenhang weniger.

4.2 Einzelevaluierung

Im Projekt MIROSLAV wird lediglich eine Bewertung eines einzelnen MT-Systems vorgenommen. Es finden keine Vergleichstests mit anderen Systemen statt.

4.3 Relevanz- und Informativübersetzung

Statt der in der Literatur üblichen Zweiteilung des Benutzerbedürfnisses in Informativ- und Reinübersetzung wird im MIROSLAV-Evaluierungskonzept eine Dreiteilung getroffen, die zusätzlich den Verwendungszweck Relevanzübersetzung umfaßt. Hierbei ist der Benutzer lediglich darin interessiert, ob ein bestimmter Text die für seinen Verwendungszweck relevante Information enthält oder nicht, d.h. ob sich ein Text inhaltlich zur Weiterverwendung wie z.B. der Datenbankaufbereitung (s.o.) eignet³.

Neben der Relevanzübersetzung werden beim MIROSLAV-Anwendungspartner GESIS-Berlin hauptsächlich Informativübersetzungen benötigt, da die russischsprachigen Originaltexte nicht in ihrer Gänze verarbeitet werden, sondern vielmehr als Ausgangsbasis für die Pflege einer breiten Palette von Informationsmedien (Printmedien, Datenbanken) dienen. Hierzu extrahiert man aus dem Original spezifische Informationen und bereitet sie gemäß dem Format des Zielmediums auf (s. Marx 1998a und 1998b).⁴

4.4 Übersetzungsqualität und -produktivität

Für den Nutzer eines automatischen MT-Systems ist nicht nur die Qualität der erzeugten Texte interessant, sondern auch, ob sich durch den Einsatz eines solchen Systems eine Arbeitserleichterung oder Produktivitätssteigerung erzielen läßt. Hierzu muß sich die Übersetzungskomponente gut in den Arbeitsprozeß eingliedern lassen, leicht zu bedienen sein, ggf. schnell arbeiten usw. Deshalb ist sowohl

1998). Hier etabliert sich zur Zeit gerade der Begriff der task-orientierten Evaluierung (s. z.B. White/Taylor 1998 und Hovy 1998), die die Bewertung eines Systems in Abhängigkeit vom angestrebten Einsatzgebiet vornimmt. Benutzerorientiert ist darüber hinaus eine Evaluierung in unserem Verständnis dann, wenn sie zusätzlich den potentiellen Benutzer des zu bewertenden Systems in Betracht zieht, also z.B. auch individuelle und psychologische Kriterien (z.B. über eine Benutzermodellierung oder konkrete Benutzertests) berücksichtigt.

³ Je nach Anwendungsdomäne sind hier natürlich weitere Verwendungszwecke denkbar. So unterscheiden bspw. White/Taylor 1998 insgesamt 6 verschiedene sogenannte Text-Handling Tasks, die von Publishing (d.i. in unserer Terminologie die Reinübersetzung) über Extraction (d.i. Informativübersetzung) bis hin zu Filtering (d.i. Relevanzübersetzung) reicht.

⁴ Darüber hinaus werden in stark begrenztem Umfang auch Übersetzungen zur Weiterverwendung benötigt, z.B. bei der Publikation eines Themenbandes.

die Qualität als auch die Produktivität des MIROSLAV-Systems Gegenstand der Evaluierung in MIROSLAV.

4.5 Globaler Evaluierungsprozeß

Die Evaluierung in MIROSLAV soll anhand eines globalen Evaluierungsprozesses geschehen, d.h. daß alle verfügbaren Informationsquellen bei der Beurteilung der Übersetzungsqualität und -produktivität heranzuziehen sind. So werden nicht nur Herstellerangaben, Handbücher u.ä. zum System konsultiert, sondern auch Benutzerbefragungen und Systemtests mit potentiellen Benutzern durchgeführt.

4.6 Black-Box-Ansatz + Lexikonfenster

Kommerzielle Produkte erlauben dem Benutzer keinen Einblick in den Systemkern der Übersetzungskomponente, nicht zuletzt um zu verhindern, daß durch Systemmanipulation Funktionseinbußen oder unerwünschte Seiteneffekte eintreten. Eine Erweiterung oder Korrektur von Lexikoneinträgen ist dagegen jedoch üblich und auch sinnvoll.

Da die MIROSLAV-Evaluierung eine benutzerorientierte sein soll, verfolgt sie eben genau den Ansatz, den ein kommerzielles System zuläßt: Black-Box-Ansatz, d.h. ohne Einsicht in den Systemkern mit Lexikonfenster zur Anpassung des Lexikons.

4.7 Häufigkeit der Evaluierung

Zur Beurteilung der Übersetzungsqualität wird zunächst eine einmalige Durchführung der geplanten Tests angesetzt. Da von Seiten der GMS der russisch-deutsche Thesaurus des IZ Bonn in das MIROSLAV-System integriert wird, erscheint ein Lexikontuning und damit ein zweiter Testdurchlauf aus momentaner Sicht nicht erforderlich.

Bezüglich der Produktivität müssen dagegen zumindest zwei Durchgänge gestartet werden, da die Arbeitseffektivität vor und nach dem Einsatz eines maschinellen Übersetzungssystems festzustellen ist. Machen die Ergebnisse dieser Tests eine vollständige Reorganisation der Arbeitsprozesse erforderlich, so ist gegebenenfalls ein dritter Durchlauf erforderlich.

4.8 Evaluierungsziele: Qualität

Gängige Kriterien zur Feststellung der Übersetzungsqualität eines Systems sind im allgemeinen die Lesbarkeit, Verständlichkeit und die semantische Treue⁵ des Zieltextes, sowie die Art und Anzahl der im Output auftretenden Fehler. Da die Lesbarkeit eines Textes nichts über dessen Adäquatheit in Bezug auf den Quelltext aussagt (ein gut lesbarer Text kann inhaltlich falsch sein), wird sie bei der MIROS-

⁵ Statt „semantische Treue“ wird in der Literatur oft auch der Begriff „Informativität“ verwendet. Wegen der besonderen Bedeutung des Terminus „Information“ in der Informatikwissenschaft, der hier eine stark pragmatische Ausrichtung hat und den Neuheitswert eines Datums für den Benutzer umschließt, soll auf den u.E. passenderen Ausdruck „semantische Treue“ zurückgegriffen werden.

LAV-Evaluierung nicht in Betracht gezogen. Um ein möglichst umfassendes Bild der Übersetzungsqualität des MIROSLAV-Systems zu gewinnen, sollen jedoch die restlichen Kriterien in die Evaluierung einbezogen werden, wie folgende Aufstellung zeigt.

- **Verständlichkeit**

Um die Verständlichkeit der von METAL übersetzten Texte beurteilen zu können, bietet sich ein Rating-Verfahren an. Das Rating-Verfahren ist die am häufigsten angewandte Methode zur Messung der Verständlichkeit maschinell übersetzter Texte. Es basiert auf der Verwendung einer Skala, nach der Texteinheiten, meist Sätze gemäß ihrer Verständlichkeit einzustufen sind. Da sich die Verwendung umfangreicher Skalen in der Vergangenheit als äußerst problematisch herausgestellt hat (s. Marx 1998a), wendet MIROSLAV eine lediglich 3-stufige Skala an mit den Meßpunkten „verständlich“, „teilweise verständlich“ und „unverständlich“. Bemessungseinheit ist jeweils ein Satz bzw. eine Phrase im Falle von Überschriften, Aufzählungen etc.

- **Semantische Treue**

Zur Bestimmung der semantischen Treue, die ein Maß für die Genauigkeit und Treue einer Übersetzung ist, wird ein sogenannter Deskriptorentest durchgeführt. Hierbei handelt es sich um ein eigens für den MIROSLAV-Kontext entwickeltes Testverfahren, das sich aus den speziellen Gegebenheiten des Projektanwendungspartners ergibt. Möglich wird dieses Verfahren durch die Tatsache, daß die GESIS-Berlin einen Großteil der recherchierten Texte für die Projektdatenbank FORIS sowie die Institutionendatenbank aufbereitet (s.o.), d.h. u.a. formale und inhaltliche Schlagwörter vergibt, um die Suche nach Dokumenten zu ermöglichen. Angelegt ist das vorgeschlagene Verfahren auf der Basis eines Vergleichstestes, bei dem die Art und Anzahl der für einen Originaltext vergebenen formalen und inhaltlichen Deskriptoren verglichen wird mit der Art und Anzahl der Deskriptoren, die anhand der maschinellen Übersetzung dieses Textes vergeben wurden. Je geringer die Unterschiede zwischen beiden Gruppen, desto größer ist die Genauigkeit und Verständlichkeit des maschinell übersetzten Textes. Die Hauptidee ist hierbei, daß die Deskriptoren jeweils für die wichtigsten, die den Textinhalt bestimmenden semantischen Einheiten vergeben werden und somit die Deskriptorenübereinstimmung zwischen Original und Übersetzung ein, wenn auch grober, Indikator für die Übereinstimmung in punkto semantische Treue gelten kann.

Außerdem macht der Deskriptorentest auch Aussagen über die Verständlichkeit eines maschinell übersetzten Textes, da nur ein (gut) verständlicher Text korrekt verschlagwortet werden kann.

- **Fehleranalyse**

Da, wie bereits beschrieben, die MIROSLAV-Evaluation keine diagnostische oder therapeutische Analyse zum Ziel hat, soll bei der Fehleranalyse der maschinellen Übersetzung eine eingeschränkte Fehlerklassifikation zur Anwendung kommen. Diese berücksichtigt lediglich diejenigen Fehlertypen, die massive Auswirkungen auf die Verständlichkeit des Textes haben. Bei der Festlegung der zu berücksichtigenden Fehlertypen haben wir uns an Rinsche 1993 sowie

an Flanagan 1994 orientiert. Demnach ergibt sich folgendes Raster, das anhand einer kleinen Textauswahl der Anwendungsdomäne auf ihre Tragbarkeit hin überprüft wurde:

Berücksichtigt werden sollen:

1. falsche Segmentierung:
Die einzelnen Satzglieder sind nicht richtig erkannt worden bzw. Einzelwörter oder Satzteile werden den falschen Satzgliedern zugeordnet.
2. falsche Satzgrenzen:
Die Satzgrenzen sind falsch gesetzt, entweder durch „Zerstückeln“ eines Satzes oder Zusammenziehen zweier aufeinanderfolgender Sätze.
3. falsche Ellipsen:
Das System interpretiert eine Struktur fälschlicherweise als Ellipse, übersetzt z.B. eine Nominalphrase als Satz mit dem Kopulaverb *byt'*.
4. Wort nicht übersetzt:
Ein Wort ist nicht im Lexikon vorhanden oder wird aus anderen Gründen (z.B. wegen phrasaler Analyse) nicht übersetzt (mit Ausnahme der nicht zu berücksichtigenden Wortklassen, s.u.).
5. falsche Wortwahl:
Ein Wort wird falsch, d.h. nicht kontextgemäß übersetzt.

Diese Hauptfehlerklassen sind nach absteigender Wichtigkeit gerankt. Da die einzelnen Fehler unterschiedliche Auswirkungen auf die Verständlichkeit und damit die Brauchbarkeit einer Übersetzung haben, erscheint eine solche Gewichtung sinnvoll, um zu realistischen Aussagen über die Übersetzungsqualität zu gelangen. Dieses Ranking hat allerdings erst vorläufigen Charakter und ist an einer größeren Textauswahl oder auch am eigentlichen Testmaterial zu überprüfen.

Neben den oben genannten, zu berücksichtigenden Fehlertypen sind mit der Negation, Modalverben und Komposita bzw. Mehrwortausdrücken noch drei weitere Felder zu nennen, die bei fehlerhafter Übersetzung u.E. massive Auswirkungen auf das Satz- und Textverständnis haben können. Allerdings traten solche Fehler in den von uns untersuchten Texten nicht auf, so daß eine Entscheidung über eine Aufnahme dieser Kategorien in die Gruppe der zu berücksichtigenden Fehler noch aussteht, bzw. erst anhand der eigentlichen Fehleranalysetests gefällt werden kann.

Nicht zu berücksichtigende Fehlerklassen sind:

- Großschreibung:
Die Groß-/Kleinschreibung hat keinen Einfluß auf die Grammatikalität und damit die Verständlichkeit eines Satzes.
- Rechtschreibung
In der überwiegenden Mehrzahl der Fälle bleibt ein Wort in der Zielsprache auch bei fehlerhafter Rechtschreibung verständlich.
- Wortstellung (d.i. Satzgliedstellung)

Eine korrekte Segmentierung (s.o.) vorausgesetzt, bleibt ein Satz auch bei fehlerhafter Wortstellung verständlich.

– Auslassungen

Auslassungen treten praktisch nicht auf. Kann ein Wort oder eine Phrase nicht übersetzt werden, so wird die Originalform in die Übersetzung übernommen.

– Flexion (Kongruenz, Tempus)

Bei falscher Flexion bleibt ein Satz nach wie vor verständlich bzw. die Satzbedeutung rekonstruierbar.

– Passiv

Auch Fehler bei der Passivbildung bzw. eine fälschlicherweise nicht erfolgte Passivierung beeinträchtigen das Satzverständnis in der von uns untersuchten Textauswahl nicht massiv.

– Funktionswörter (auch Präpositionen und Konjunktionen)

Außer bei den lokalen Präpositionen kann die Bedeutung dieser Wortart aus dem Kontext erschlossen werden. Für die Konjunktionen scheint zu gelten, daß sich das logische Verhältnis zweier verknüpfter Sätze mehr aus dem semantischen Gehalt der Einzelsätze zueinander ergibt als aus der Bedeutung der Konjunktion selbst. Zudem traten in der von uns untersuchten Textauswahl keine Fehler bei der Übersetzung von Konjunktionen auf, so daß es vertretbar scheint, diese Fehlerart aufgrund mangelnder Relevanz unberücksichtigt zu lassen.

– Pronomen

Die Bedeutung der Pronomen ist meistens aus dem Kontext ersichtlich bzw. rekonstruierbar.

4.9 Evaluierungsziele: Produktivität

Die für die Produktivität eines automatischen Übersetzungssystems relevanten Faktoren können eingeteilt werden in die Gruppen Hardware, Software und anwenderspezifische Kriterien. Wie diese im Rahmen von MIROSLAV zu behandeln sind, soll im folgenden vorgestellt werden.

- Hardware

Hardwarefaktoren, die das Laufzeitverhalten eines Systems und somit seine Produktivität beeinflussen, sind der Prozessor, Taktfrequenz, benötigter Festplatten- und Arbeitsspeicher, Netzwerkfähigkeit sowie die Druckgeschwindigkeit des Druckers. Diese Daten sind anhand der Herstellerangaben leicht meßbar. Weiterhin entscheidend ist die Stabilität des Gesamtsystems, die im Rahmen der Benutzertests zur Produktivität der Software (s.u.).

- Software

Die Produktivität einer Software ist von einer Vielzahl von Einzelfaktoren abhängig, die jeweils einen mehr oder weniger großen Einfluß auf die Qualität und Quantität des Outputs nehmen. Folgende Tabelle, die sich bei der Einteilung an der ISO-Norm 9126 zu Qualitätscharakteristika von Softwareprodukten orientiert, gibt eine Aufstellung der im Rahmen der MIROSLAV-Evaluierung zu berücksichtigenden Kriterien.

Funktionalität	Verlässlichkeit	Benutzbarkeit	Effizienz	Portabilität
Texthandling (Import- Einbindung in Arbeitsumgebung Netzwerkanbindung Übersetzungsarchive (Translation Lexikonhandling (Kodiertool) Einstellung von Sachgebieten Angabe von Übersetzungseinheiten Postediting-Unterstützung interaktives Übersetzen Voreinstellungen zur Übersetzung wie Anrede)	Robustheit (Feh-	Mensch-Maschine-Interaktion (Bild- , Lernbarkeit, Operationalität) ⁶ Inkonsistenzvermeidung bei Lexiko- Qualität der Systemdokumentation	Geschwindigkeit	Installierbarkeit Plattformunabhän-

⁶ Kriterien zur Softwareergonomie, wie sie bspw. in ISO 9126 genannt werden, sind aufgrund ihrer mangelnden Operationalisierbarkeit äußerst umstritten. Auf diese Diskussion kann allerdings im Rahmen der MIROSLAV-Evaluierung nicht eingegangen werden; es sei hier auf Krause/Womser-Hacker 1997 verwiesen. Das Spektrum der Mensch-Maschine-Interaktion wird bei der Produktivitätsanalyse von METAL anhand von Benutzertests untersucht.

Auf die Überprüfung der Wartbarkeit, zu der Kriterien wie Einfachheit der Fehleranalyse, Erweiterbarkeit u.ä. zählen, kann wegen der bereits erwähnten fehlenden diagnostischen Sichtweise der MIROSLAV-Evaluierung verzichtet werden.

Für die Vielzahl der softwareergonomischen Kriterien, die bei der Bewertung der Produktivität eine Rolle spielen, ist kein einheitliches Testdesign möglich. Hier sind je nach Art und Wichtigkeit des zu evaluierenden Bereichs unterschiedliche Verfahren anzuwenden wie z.B. direkte Evaluierung durch Auflistung der im System angebotenen Funktionen, analytische Bewertung der Benutzbarkeit durch eingehende Sichtung des Systems sowie Benutzertests mit Videoüberwachung und vorgegebenem Aufgabenset u.ä. (s. Marx 1998b).

- Anwenderspezifische Kriterien
Neben den Hardware- und Softwarekriterien gibt es noch eine Reihe systemexterner Faktoren, die die Produktivität eines maschinellen Übersetzungssystems beeinflussen. Dazu sind zu zählen:

Workflow

Wo und wie kann das Übersetzungssystem in den aktuellen Arbeitsablauf der Anwendungsdomäne eingesetzt werden? Sind durch den Einsatz eines maschinellen MT-Systems Produktivitätssteigerungen zu erwarten oder eingetreten? Hierbei sind nicht nur Quantitätssteigerungen beim Output ausschlaggebend, sondern auch Verbesserungen, die der Stabilität des gesamten Workflow, wie z.B. Dezentralisierungen dienen.

Textart

Die Produktivität eines Übersetzungssystems ist immer abhängig von der Art der zu übersetzenden Texte, da die einzelnen Textsorten von recht unterschiedlicher Komplexität sind. Daher ist die Angabe der bearbeitenden Textart unerlässlich. Da in MIROSLAV keine elaborierte Texttypologie geleistet werden kann, soll die Angabe einiger wesentlicher Eckdaten genügen:

- äußere Form: Ist das Dokument maschinenlesbar oder handelt es sich um einen Schreibmaschinentext?
- formale Kriterien: Wie hoch ist der Anteil an Graphiken, Tabellen, Aufzählungen, Fließtext etc.?
- grammatische Komplexität: Wie groß ist die durchschnittliche Satzlänge? Enthält der Text viele Einbettungen, komplexe Nominalphrasen, (Genitiv-)Attribute u.ä.?

Benutzer motivation

Die Motivation zur Nutzung eines maschinellen Übersetzungssystems beim MIROSLAV-Anwendungspartner kann nach Gesprächen vor Ort als sehr hoch eingestuft werden. Die mit der Bearbeitung russischer Texte betrauten Mitarbeiterinnen sind sehr am Einsatz einer solchen Komponente interessiert und stehen ihr positiv gegenüber. Sie erwarten sich dadurch eine Arbeitsentlastung sowie eine Dezentralisierung bei der Arbeitsvergabe. Näheres ist durch einen Fragebogen zu ermitteln.

System- und Hardwarekenntnisse

Die System- und Hardwarekenntnisse beeinflussen die Produktivität eines Systems hauptsächlich insofern, als sie entscheidend sind für den Umgang der Benutzer mit Störsituationen. Sie sind durch einen Fragebogen zu ermitteln, in den auch Fragen zur Benutzer motivation eingehen sollen.

5 Zusammenfassende Übersicht über die MIROSLAV-Testverfahren

Nachfolgende Aufstellung gibt nochmals einen zusammenfassenden Überblick über die im Projekt MIROSLAV geplanten Evaluierungstests, getrennt nach Tests mit Benutzerpartizipation, sogenannte Benutzertests und denjenigen Testverfahren, die von MIROSLAV-Projektmitarbeitern durchzuführen sind.

Evaluation durch MIROSLAV-Projektmitarbeiter:

- Fehleranalyse
- Hardware- und Funktionalitätscheck
- Softwareergonomieanalyse
- Effizienztest
- Portabilitätstest

Benutzertests:

- Rating-Test
- Deskriptorentest
- Softwareergonomietest
- Workflowtest
- Umfrage zur Benutzer motivation

6 Ausblick

Mit dem Erscheinen der ersten PC-Version T1 für Deutsch-Russisch von Langenscheidt wird im Herbst diesen Jahres gerechnet. Sobald die Software vorliegt, kann mit der Testphase der MIROSLAV-Evaluierung begonnen werden. Bis Winter 1998/Frühjahr 1999 kann mit ersten Ergebnissen gerechnet werden, die zeigen, ob sich das theoretische Grundkonzept in der Praxis bewährt.

7 Literatur

[Bourbeau 1990]

Bourbeau, L. (1990): Elaboration et mise au point d'une méthodologie d'évaluation linguistique de systèmes de traduction assistée par ordinateur. Rapport final. Secrétariat d'État du Canada.

[Flanagan 1994]

Flanagan, M. (1994): Error Classification for MT Evaluation. AMTA 94.

[Hovy 1998]

Das Konzept einer strikt benutzerorientierten Evaluierung maschineller Übersetzungssysteme

Hovy, E. (1998): Creating Useful Metrics for Evaluating Machine Translation. Vortrag auf der First International Conference on Language Resources & Evaluation. Granada.

[Krause/Womser-Hacker 1997]

Krause, J.; Womser-Hacker, C. (1997): Benutzerfreundlichkeit durch graphische Benutzungsoberflächen und die Integration von Vagheit. In: Krause, J.; Womser-Hacker, C. (eds.): Vages Information Retrieval und graphische Benutzungsoberflächen: Beispiel Werkstoffinformation. Schriften zur Informationswissenschaft 28. Konstanz.

[Marx 1998a]

Marx, J. (1998a): Vorarbeiten zur Evaluierung der Übersetzungsqualität und -produktivität von METAL (russisch-deutsch). In: MIROSLAV/R-Bericht 5/3. Regensburg.

[Marx 1998b]

Marx, J. (1998b): Typologie der Evaluierung im Projekt MIROSLAV. Evaluierungskonzept, gewählte Testverfahren und -design. MIROSLAV/R-Bericht 6/3. Regensburg.

[Marx/Mutschke/Schommler]

Marx, J.; Mutschke, P.; Schommler, M. (1995): Möglichkeiten der intelligenten Integration heterogener Datenbestände - Das Projekt GESINE; Dezember 1995; ISSN: 1431-6943; Herausgeber: Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI).
<http://www.bonn.iz-soz.de/publications/series/working-papers/ab2/ab203.htm>

[Rinsche 1993]

Rinsche, A. (1993): Evaluationsverfahren für maschinelle Übersetzungssysteme: zur Methodik und experimentellen Praxis. Technical Report, Kommission der Europäischen Gemeinschaft, Bericht EUR 14766 DE.

[Rubio/Gallardo/Castro/Tejada 1998]

Rubio, A.; Gallardo, N.; Castro, R.; Tejada, A. (1998) (eds.): First International Conference on Language Resources & Evaluation. Proceedings. Granada.

[Van Slype 1979]

Van Slype, G. (1979): Critical study of methods for evaluating the quality of machine translation. Final report. Bruxelles: Bureau Marcel van Dijk.

[Vasconcello 1988]

Vasconcellos, M. (1988): Factors in the Evaluation of MT: Formal vs. Functional Approaches. In: Vasconcellos, M (ed.) (1988): Technology as Translation Strategy. Binghamton: State University of New York. pp. 203 - 213.

[Way 1994]

Way, A. (1994): Developer-Oriented Evaluation of MT Systems. In: Falkedal, K. (ed.): Proceedings of the Evaluators' Forum, 1991, Les Rasses, Vaud, Switzerland. Genf: ISSCO. pp. 237 - 244.

[White/Taylor 1998]

White, J.S.; Taylor, K.B. (1998): A Task-Oriented Evaluation Metric for Machine Translation. In: Rubio, A.; Gallardo, N.; Castro, R.; Tejada, A. (eds.): First International Conference on Language Resources & Evaluation. Proceedings. Granada.