BMBF

Verbmobil
Verbundvorhaben

Übersetzungsstrategien, Bewertung und Kontrolle für VERBMOBIL

Schlußbericht der Arbeitspakete 13.2, 13.3, 13.8

2., leicht überarbeitete Version

Christa Hauenschild
Susanne Heizmann
Susanne Petzolt
Birte Prahl

Universität Hildesheim

Vm
Report 203
August 1997

August 1997

Christa Hauenschild
Susanne Heizmann
Susanne Petzolt
Birte Prahl

Universität Hildesheim
Institut für Angewandte Sprachwissenschaft
Computerlinguistik
Marienburger Platz 22
D- 31141 Hildesheim

Tel.: +49 5121 883 - 342
e-mail: chau@cl.uni-hildesheim.de

Gehört zu den Antragsabschnitten: 13.2 Übersetzungsstrategien I und
13.3 Übersetzungsstrategien II und
13.8 Kontrolle und Bewertung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 P 5 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autorinnen.

Zusammenfassung

Der vorliegende Bericht präsentiert die wichtigsten Ergebnisse der Arbeitspakete 'Übersetzungsstrategien I' und 'Übersetzungsstrategien II' (13.2, 13.3) und 'Bewertung und Kontrolle' (13.8) des Verbundprojektes VERBMOBIL. Er dokumentiert die Arbeit des Teilvorhabens von VERBMOBIL in Hildesheim. Als Hauptergebnisse liegen für die Arbeitspakete 'Übersetzungsstrategien' eine Beschreibung menschlicher Strategien, die für VERBMOBIL nutzbar sind, und ein Konzept der variablen Analysetiefe vor sowie für das Arbeitspaket 'Bewertung und Kontrolle' eine Kriterienaufstellung zur Bewertung von VERBMOBIL und Konzepte zur Kontrolle durch die NutzerInnen. Wir gehen außerdem auf das Konzept des Translationsziels ein, das sich als zentral für unsere Arbeiten und für VERBMOBIL herausgestellt hat.

Inhaltsverzeichnis

1 Einleitung

In diesem Schlußbericht werden die Ergebnisse der Arbeitspakete 13.2, 13.3 und 13.8 zusammengefaßt.[1] Für Details zu den jeweiligen Einzelaspekten werden wir auf die entsprechenden VM-Memos und -Reports verweisen, wo auch die genauen Angaben zur verwendeten Fachliteratur zu finden sind.
Die Untersuchungen des Hildesheimer Teilvorhabens hatten den menschlichen Translationsprozeß[2] als Ausgangspunkt, d.h. es wurden Verfahren und Kriterien der Humantranslation im Hinblick auf Verwertbarkeit für VERBMOBIL erforscht.
Das gewählte Verfahren umfaßte zwei Schritte:

- Da die wenigen Vorarbeiten im Bereich prozeßorientierter translationswissenschaftlicher Forschung für die MT (Machine Translation) nur begrenzt nutzbar sind, mußten unsere Arbeiten explorativer Natur sein. Es konnten daher im ersten Schritt nur Arbeitshypothesen über menschliche Strategien und Bewertungen generiert werden, die allerdings einen hohen Grad an Plausibilität haben sollten.

- Da sich der maschinelle Dolmetschprozeß in vieler Hinsicht deutlich vom menschlichen unterscheidet, mußten im zweiten Schritt aus den Ausgangshypothesen wiederum Hypothesen über Konzepte und Strategien für VERBMOBIL abgeleitet werden.

Es ist somit klar, daß nicht jede Ausgangshypothese in eine plausible Strategie für VERBMOBIL einmünden konnte. Wir wollen im folgenden zeigen, welche unserer Ausgangshypothesen wir für vielversprechend für VERBMOBIL halten und wie die Umsetzung aussieht bzw. aussehen könnte.
Der Translationsprozeß wird von zwei Seiten her beschrieben: Zum einen untersuchen wir menschliche Strategien zur Lösung typischer Translationsprobleme wie z.B. zum Umgang mit Wissensdefiziten (Arbeitsbereich 'Übersetzungsstrategien'). Zum anderen entwickeln wir Kriterien der Bewertung maschineller Translationsleistungen aus Nutzersicht sowie Modelle zur Kontrolle maschineller Leistungen durch die NutzerInnen (Arbeitspaket 'Bewertung und Kontrolle').
Als Hauptergebnis für das Arbeitspaket 'Übersetzungsstrategien' stellen wir ein Konzept der Variablen Analysetiefe vor. Der Weg dorthin sowie die wichtigsten Teilergebnisse werden in Abschnitt 2 beschrieben.
Für das Arbeitspaket 'Bewertung und Kontrolle' wird eine Liste von Bewertungskriterien aus Nutzersicht erläutert sowie einige Vorschläge zur Kontrolle durch die NutzerInnen (Abschnitt 3).

[1] Wir verwenden im folgenden die Abkürzungen AP = Arbeitspaket, TP = Teilprojekt, VM = VERBMOBIL.

[2] Wir verstehen den Begriff 'Translation' als Oberbegriff zu 'Übersetzen' und 'Dolmetschen'.

Im Laufe unserer Arbeiten hat sich herausgestellt, daß ein operationalisierbares Kriterium für die Adäquatheit von menschlichen und maschinellen Translationen ein absolutes Desideratum ist, da die bisher in der maschinellen Übersetzung verwendeten Konzepte für VERBMOBIL offensichtlich unzureichend sind. Daher wurde gemeinsam mit den PartnerInnen in Hamburg ('Dolmetschstrategien') und Berlin ('Semantische Auswertung') das Konzept des **Translationsziels** entwickelt, das in Abhängigkeit von den Gegebenheiten der Kommunikationssituation Adäquatheitskriterien für die Translation zur Verfügung stellt. Dieses Konzept wird zunächst in zwei Versionen vorgestellt und dann in seiner Bedeutung für VERBMOBIL dargestellt (Abschnitt 4). In einem Fazit (Abschnitt 5) wird schließlich eine allgemeine Einschätzung der Ergebnisse gegeben.

2 Übersetzungsstrategien

2.1 Fragestellung

Das Arbeitspaket 'Übersetzungsstrategien' untersucht die Strategien, die menschliche Translatoren anwenden, um mit Wissenslücken im Translationsprozeß umzugehen. Auch Human-Translatoren haben nicht immer alle translationsrelevante Information zur Verfügung, um die anstehenden Entscheidungen im Translationsprozeß treffen zu können. Da der Translationsprozeß grundlegend ein Entscheidungsprozeß ist, in dem praktisch nie *alle* Informationen unverändert wiedergegeben werden können, müssen Human-Translatoren ständig - bewußt oder unbewußt - Entscheidungen treffen, die die Relevanz von Informationen betreffen. Diese Entscheidungen werden häufig auf der Grundlage von unsicherem Wissen getroffen (z.B. aufgrund unzureichender Kenntnisse über die Kommunikationssituation und deren Hintergründe).
Interessant ist, welche Strategien Human-Translatoren anwenden, um dennoch Entscheidungen zu treffen, die eine adäquate Translation ermöglichen. Auf der Grundlage welchen Wissens entscheiden sie über die Relevanz von fehlender und vorhandener Information?
Um diese Fragen zu beantworten, haben wir eine Reihe von experimentellen Studien durchgeführt, auf die wir in den folgenden Abschnitten bei Bedarf genauer eingehen werden.

2.2 Translationsprobleme - Translationsstrategien

Translationsprobleme standen am Anfang unserer Untersuchungen, da wir unserer Arbeit eine problembasierte Definition des Strategiebegriffs zugrunde legten. Es sollte zunächst beschrieben werden, was für einen Human-Translator überhaupt problematisch ist, um daraufhin untersuchen zu können, welche Strategien er einsetzt, um mit diesen Problemen umzugehen.

Für eine vorläufige Klassifikation der Translationsprobleme mußten zentrale Begriffe neu definiert werden, da bisher keine für unsere Zwecke brauchbare Definition von Translationsproblemen existierte. Die relevanten Konzepte, die für unsere Untersuchungen grundlegend sind, haben wir in einem ersten Ansatz in [VM-Memo4] beschrieben.
Der in der Translationswissenschaft verwendete Begriff des Translationsproblems mußte dabei modifiziert werden. Üblicherweise wird die Bewußtheit von Translationsproblemen als Kriterium verwendet, was auf VERBMOBIL als maschinelles Dolmetschsystem nicht anwendbar ist. Für unsere Zwecke mußten objektive Kriterien gefunden werden, die nicht von der Frage der Bewußtheit oder Unbewußtheit von Translationsproblemen abhängen. Wir unterscheiden zwischen **potentiellen** und **aktuellen** Translationsproblemen, wobei es nur die aktuellen Translationsprobleme sind, die für VERBMOBIL relevant werden.
Nach unserer Auffassung ist das grundlegende Kennzeichen des aktuellen Translationsproblems ein Wissensdefizit zu einem bestimmten Zeitpunkt im Translationsprozeß [VM-Memo4]. Nicht jedes Wissensdefizit führt zu einem aktuellen Translationsproblem, aber jedes aktuelle Translationsproblem beinhaltet ein Wissensdefizit. **Ein Translationsproblem tritt demnach auf, wenn für eine zu treffende Entscheidung im Translationsprozeß zu einem bestimmten Zeitpunkt Information fehlt.**
Bei Wissensdefiziten haben Human-Translatoren grundsätzlich zwei Möglichkeiten (s. auch [VM-Memo4]):

1. Sie entscheiden, daß die fehlende Information wenig relevant ist, und produzieren einen zielsprachlichen Text, der weniger Information enthält als das Original, ohne dabei das Erreichen des Kommunikationsziels zu gefährden. Damit wenden sie eine **Reduktions-Strategie** an (z.B. in Form einer Generalisierung, s. auch [VM-Memo34]).

2. Sie entscheiden sich für eine **Achievement-Strategie**, d.h. sie beurteilen die fehlende Information als relevant und bemühen sich, zusätzliche Information zu beschaffen (z.B. durch Nachfragen beim Autor).

Die beiden Typen von Strategien sind in Anlehnung an Strategien des Zweitspracherwerbs definiert.

2.3 Menschliche Desambiguierungsstrategien

Nach dieser grundsätzlichen Klassifikation menschlicher Translationsstrategien untersuchten wir in Zusammenarbeit mit TP11 ‘Semantische Auswertung’ ein relevantes Fragment des VERBMOBIL-Korpus mit repräsentativen Desambiguierungsaufgaben im Hinblick auf menschliche Lösungsansätze.

Dabei wurde schnell deutlich, daß im Dolmetschprozeß vor allem Reduktionsphänomene eine wichtige Rolle spielen, so daß wir eine genauere Untersuchung von Reduktionen im für uns relevanten VERBMOBIL-Korpus (TP-13 Gespräche, VM-Memo 18) durchgeführt haben.
Im folgenden ist ein Beispiel für Reduktion angeführt, das dem TP13-Korpus entnommen wurde:[3]

(36)	MAY	⟨ 2s⟩ Ähm ⟨ 2s⟩ ich würde vorschlagen also möglich wär s bei mir, daß man ⟨P⟩ den Mittwoch, den dritten ⟨P⟩ November ⟨P⟩ nimmt. ⟨Stöhnen⟩
(37)	MAY	⟨ ?s⟩ Da hätt ich Zeit ⟨P⟩ äh' vormittags Zeit.
(38)	DEF	Mister Mayer proposes ⟨P⟩ November third in the morning.

Aus den Daten wurde klar, daß zwei grundlegende Formen von Reduktion zu unterscheiden waren (s. [VM-Memo34]):

- Reduktion als spezielle Problemlösestrategie: **strategische Reduktion**
- Reduktion als ein dem Translationsprozeß - und vor allem dem Dolmetschprozeß - inhärentes Verfahren (z.B. Reduktion von Häsitationsphänomenen): **nicht-strategische Reduktion**

Strategische Reduktion wird als Strategie zum Lösen von Translationsproblemen angewandt, d.h. daß bei Wissenslücken problematische Ausdrücke abgeschwächt transferiert oder sogar weggelassen werden können, wenn sie als wenig relevant beurteilt werden.
Mit der **nicht-strategischen Reduktion** wird ein dem Translationsprozeß inhärentes Phänomen angesprochen: Da eine Translation aufgrund der bestehenden Divergenzen zwischen ausgangssprachlichem und zielsprachlichem Sprachsystem notwendigerweise auf Satz- bzw. Äußerungsebene bedeutungsabschwächend oder bedeutungsverstärkend sein **muß**, gehören Reduktionsphänomene zu den regulären Bestandteilen des Translationsprozesses. Diese Bedeutungsabschwächung auf Mikroebene, d.h. auf der Ebene einzelner Sätze oder Ausdrücke, hat auf die Erreichung des kommunikativen Ziels keinen Einfluß, da die entsprechende Information nicht zur intendierten Interpretation der Äußerung gehört.[4]
Eine erste Analyse des relevanten Korpus (TP 13 Gespräche) in bezug auf Reduktionsphänomene wurde gemäß den Verarbeitungsebenen (Systemkomponenten) von VERBMOBIL durchgeführt (beschrieben in [VM-Memo34]).

[3] [VM-Memo18], Gespräch 11, MAY ist hier der deutsche Gesprächspartner, DEF ein Dolmetscher mit Englisch als Fremdsprache.

[4] Das Konzept der "intendierten Interpretation" wurde genauer ausgearbeitet in der Dissertation von Birte Schmitz (TU Berlin, TP 11), die demnächst als VM-Report vorliegen wird.

Als Beispiel für Reduktion auf der Ebene der Signalerkennung haben wir etwa die folgenden Formen der Reduktion unterschieden:[5]

1. Unterdrückung von Störgeräuschen

   1a) Störgeräusche vom Sprecher verursacht

   z.B. ⟨äh⟩

   in: 'Fraglich ⟨!frachlich⟩ ist , ob ⟨P⟩ ⟨äh⟩ Fronleichnam ⟨P⟩ ⟨äh⟩ in ⟨P⟩ Hessen ,'

   1b) Störgeräusche von außen

   z.B. ⟨#Kugelschreiber klappert⟩

2. Weglassen oder Kürzen von Pausen

   z.B. ⟨P⟩

   in: 'ob ⟨P⟩ ⟨äh⟩ Fronleichnam ⟨P⟩ ⟨äh⟩ in ⟨P⟩ Hessen'

Die Auswertung weiterer Studien zur Untersuchung menschlicher Desambiguierungsstrategien (Studie mit Studierenden und professionellen ÜbersetzerInnen [VM-Report197], Studie mit professionellen DolmetscherInnen [VM-Memo 73] und Studie zur Hypothesenrevision mit professionellen ÜbersetzerInnen und der Methode des Lauten Denkens [VM-Memo 97]) hat zu dem Ergebnis geführt, daß neben der Reduktion folgende Strategien eingesetzt werden:

- **ambiguitätserhaltende** (und evtl. sogar ambiguitätssteigernde) **Translation** (z.B. "Guten Tag" → "Hello")[6]
- Orientierung an den Vorgängeräußerungen des Kommunikationspartners, der gerade ihr Adressat ist (**Prinzip der Turn-übergreifenden Kohärenz** - z.B. "Guten Tag" → "Hello" gefolgt von: "Hello" → "Guten Tag")
- Verwendung von **Standardannahmen über stereotype Dialogverläufe** (z.B. übliche Bestandteile einer Terminabsprache - "a specific time" → "Termin")
- Verwendung von **Standardannahmen aus dem Alltagswissen** (übliche Zeitpunkte für geschäftliche Treffen).

[5] Quelle: [VM-Memo18], Gespräch 12.

[6] Es handelt sich hier insofern um eine anbiguitätssteigernde Wiedergabe, als "Hello" den möglichen Referenz-Zeitraum noch weniger einschränkt als "Guten Tag"; bei ambiguitätssteigernder Übersetzung liegt nach unserem Verständnis auch Reduktion vor.

Diese Punkte werden in [VM-Memo121] genauer ausgeführt, insbesondere die Strategie der Herstellung von Turn-übergreifender Kohärenz zur Auflösung verschränkter Ambiguitäten.
Alle diese Strategien dienen einem möglichst reibungslosen Gesprächsablauf und der Aufrechterhaltung eines harmonischen Gesprächsklimas (zu den Konsequenzen für VERBMOBIL s. Abschnitt 2.6).

2.4 Defaults und Hypothesenrevision

Der gesamte menschliche Translationsprozeß ist stark von Standardannahmen geprägt. Diese Human Defaults[7] sind auf allen Ebenen des Translationsprozesses und in unterschiedlichen Formen vorhanden. Standardannahmen sind vor allem deswegen von grundlegender Bedeutung, weil **Entscheidungen im Translationsprozeß oft auf unsicherem Wissen basieren**. Die Formalisierung von unsicherem Wissen gewinnt auch in Modellen des menschlichen Sprachverarbeitungsprozesses zunehmend an Bedeutung, so daß eine genauere Untersuchung der einzelnen Wissensquellen, die mit den Human Defaults verbunden sind, für die Modellierung von Translationsleistungen grundlegend ist.
Als eine vorläufige Definition von Human Defaults im Translationsprozeß stellen wir vor:
Als Human Defaults bezeichnen wir subjektive und intersubjektive Präferenzen bei der Lösung von Translationsproblemen, die aufgrund mangelnder übersetzungsrelevanter Information auftreten.
Zur genaueren Analyse menschlicher Standardannahmen wurden zunächst zwei explorative Studien mit stark heterogenen Versuchspersonen-Gruppen durchgeführt. Einzelheiten sind in [VM-Memo54] und [VM-Memo58] aufgeführt.
Aufbauend auf den Ergebnissen dieser Studien wurde ein Experiment konzipiert, bei dem ein modifiziertes VERBMOBIL-Gespräch auf Tonband aufgezeichnet und von professionellen DolmetscherInnen verdolmetscht wurde. Bei den Vorstudien hatten sich starke Unterschiede hinsichtlich der verwendeten Translationsstrategien bei LernerInnen und ExpertInnen herausgestellt: Während die LernerInnen dazu tendieren, auch sehr unsichere Annahmen zur Lösung der Translationsprobleme zu nutzen, waren die ExpertInnen häufig in der Lage, durch ambiguitätserhaltende Übersetzungen die Probleme zu umgehen.
Die bisherigen Untersuchungen im Bereich der Human Defaults machten deutlich, daß menschliche Standardannahmen eine Translation bei ungenügender Information ermöglichen können, wobei eine tiefere Analyse nicht vorgenommen wird. Es wird eine Standardtranslation verwendet, die mit mehr oder weniger großer Wahrscheinlichkeit korrekt ist. Erste Hinweise ergaben sich für eine starke Verwendung von Standardannahmen im Bereich:

[7] Der Begriff **default** leitet sich von dem in der KI verwendeten ab.

- Orts- und Zeitangaben
- deiktische Ausdrücke (s. [VM-Memo58], [VM-Memo121])

Im folgenden wird ein Beispiel für die Verwendung von Standardannahmen im Bereich von Orts- und Zeitangaben gegeben, das dem [VM-Memo58] entnommen wurde:

Der zu übersetzende Satz lautete:
• Oh, let's see. Yes, I'm free until half past one on Wednesday the third (Pause) November.
Eine Versuchsperson[8] gab zu ihrer Übersetzung
• Oh, mal sehen. Ja, am Mittwoch, dem 3. November habe ich bis 13.30 Zeit.
folgenden Kommentar:
• "sie meint wohl tagsüber mit halb zwei, also 13:30"

Wir können damit festhalten, daß Human-Translatoren offensichtlich unterschiedlich tief analysieren. Die bisherigen Daten deuten an, daß menschliche ÜbersetzerInnen und DolmetscherInnen sehr schnell die Relevanz einzelner Ausdrücke bewerten können, um somit eine geeignete Translationsstrategie auszuwählen. Hier liegt offensichtlich ein großes Problem für die Übertragung auf eine Dolmetschmaschine: der Begriff der Relevanz muß auf irgendeine Weise operationalisiert werden. Das versuchen wir mit Hilfe des Konzeptes **Translationsziel**, das in Abschnitt 4 entwickelt wird.
Speziell für die Untersuchung menschlicher Verfahren zur Hypothesenrevision wurde eine weitere Studie konzipiert und mit Hilfe erfahrener ÜbersetzerInnen und DolmetscherInnen durchgeführt. Eine genaue Beschreibung findet sich in [VM-Memo97]. Aus methodischen Gründen wurde ein schriftlich fixierter Dialog als Grundlage genommen, der turn-by-turn präsentiert wurde und von den Versuchspersonen übersetzt werden sollte. Die Versuchspersonen wurden ausdrücklich auf die Möglichkeit des Zurückblätterns hingewiesen, nur das Vorblättern war nicht gestattet. Der Dialog (eine Terminabsprache) beinhaltete eine konstruierte Schwierigkeit, so daß die Versuchspersonen (VPn) am Ende des Dialoges auf einen Widerspruch stoßen mußten, wenn sie ihn aufmerksam verfolgt hatten. Das Versuchsdesign wurde anhand einer Pilotstudie getestet. Mittels der Methode des Lauten Denkens wollten wir Einblicke in die Gedankengänge bekommen, die durch diesen Widerspruch bei den Versuchspersonen ausgelöst werden.
Als Ergebnis hofften wir, einige Kriterien zu erhalten, die bei einer Hypothesenrevision eine Rolle spielen.
Es ergaben sich verschiedenartige Vorgehensweisen der VPn bei Widersprüchen:

[8] [VM-Memo58], Satz9, Vp7.

1. Vages Übersetzen (Reduktion): Widerspruch wird nicht aufgelöst, sondern durch vages Übersetzen vermieden.

2. Willkürliche Entscheidung für eine der konkurrierenden Hypothesen; dies wird von den VPn nicht als problematisch empfunden, da sie ihr übersetzerisches Handeln als vorläufig begreifen (Hypothesenrevision als Möglichkeit)

3. Rückgriff auf Meta-Annahmen: Widersprüche werden mit nicht wohlgeformtem Text begründet und müssen deshalb nicht zwangsläufig auflösbar sein. Defekte werden irgendwie "ausgebügelt" oder in den zielsprachlichen Text übernommen.

   Beispiel: VP2/1: "[...] daß derjenige, der das gesagt hat, selber Deutscher ist, so daß ich jetzt auch mit Interferenzen rechnen bzw. auch mit falschem Englisch rechnen muß."

Insgesamt ergaben die Untersuchungen zu Defaults und Hypothesenrevision ein sehr heterogenes Bild: Human-TranslatorInnen neigen einerseits dazu, Standardannahmen auf relativ "abenteuerliche" Weise zu verwenden; andererseits vermeiden sie offensichtlich um jeden Preis, Hypothesen zu revidieren - auch um den Preis einer inkonsistenten Übersetzung. Hier wären weitere Untersuchungen erforderlich.

2.5 Das Konzept der Variablen Analysetiefe

Unsere Untersuchungen im Bereich der Desambiguierungsstrategien im allgemeinen und der Reduktionsstrategien im besonderen haben gezeigt, daß manche problematische Ausdrücke nicht weiter analysiert werden müssen und im Translat ggf. sogar ganz wegfallen können, wenn sie für eine adäquate Translation nicht relevant sind. Die Untersuchungen von real gedolmetschten Terminabsprachen zeigten, daß Verdolmetschungen stark von Reduktionsphänomenen geprägt sind. Abbildung 1 soll eine Vorstellung davon geben, wie die zugehörigen Verfahren und Strategien in den Translationsprozeß eingebunden sind.
Zunächst wird versucht, mit Verfahren für den "Normalfall" eine Translation zu finden (als Translationseinheit stellen wir uns in erster Näherung eine Äußerung vor[9]). Der Input wird daraufhin überprüft, welche Möglichkeiten für die Wiedergabe es in der Zielsprache grundsätzlich gibt; dann wird anhand eines geeigneten Adäquatheitskriteriums (zu dessen Operationalisierung s. Abschnitt 4) versucht, eine Entscheidung zwischen den verschiedenen Möglichkeiten zu treffen. Dabei kann *nicht-strategische Reduktion* eine wichtige Rolle spielen (z.B. Tilgung von Häsitationsphänomenen). Unter günstigen Umständen kann auf dieser Grundlage eine Lösung gefunden werden.

[9]Wir sind uns allerdings des damit implizierten Segmentierungsproblems durchaus bewußt.

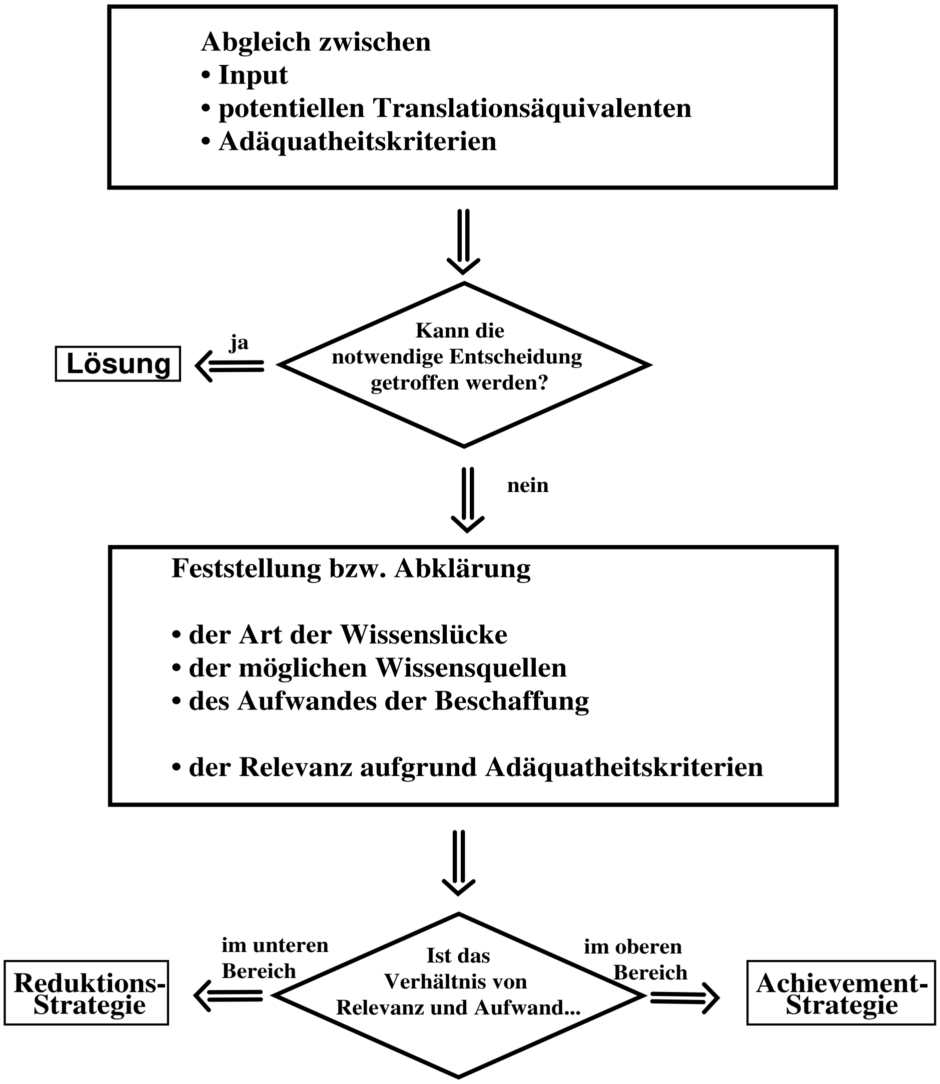

Abbildung 1: Modell der Variablen Analysetiefe

Kann die Entscheidung nicht getroffen werden, wird zunächst einmal festgestellt, welche Art von Information fehlt, wo diese Information ggf. beschafft werden kann und wie groß voraussichtlich der Aufwand dafür ist. Außerdem wird die Relevanz der fehlenden Information abgeschätzt, wieder auf der Grundlage des Adäquatheitskriteriums. Diese beiden Abschätzungen werden in eine Beziehung gesetzt, so daß bei relativ niedriger Relevanz und relativ hohem Aufwand eine Reduktions-Strategie angewandt wird, im Komplementärfall eine Achievement-Strategie (eine Achievement-Strategie in VERBMOBIL könnte das Anstoßen eines Klärungsdialogs sein).
In der ersten Version des Modells der Variablen Analysetiefe (s. [VM-Report54]) hatten wir noch "Default-Strategien" als dritte Möglichkeit zur Lösung von Translationsproblemen vorgesehen. Dies erscheint uns inzwischen nicht mehr adäquat weil die Verwendung von Defaults offensichtlich in allen Phasen des Translationsrozesses eine Rolle spielt. Aufgrund der bisher durchgeführten Studien ist es noch nicht gelungen, hier eine brauchbare Systematik zu entwicklen (s. auch Abschnitt 2.4), zumal die menschlichen Vorbilder in diesem Bereich z.T. geradezu bedenklich sind, was möglicherweise auch ein Artefakt der Versuchssituation ist. Vermutlich müßten bei der Verwendung von Defaults - ähnlich wie bei der Reduktion - auch strategische und nicht-strategische Fälle unterschieden werden. Systematisch könnte man die Default-Strategien auch als eine Klasse von Reduktions-Strategien einordnen, da sie Abstriche vom ursprünglichen Ziel im Hinblick auf die **Sicherheit der Information** machen. Hier wären weitere Untersuchungen erforderlich.

2.6 Schlußfolgerungen für VERBMOBIL

Für VERBMOBIL kann man aus den Untersuchungen zu den Übersetzungsstrategien u.E. folgende Schlußfolgerungen ziehen:

- Das wichtigste Ergebnis scheint uns das Prinzip der **Reduktion** zu sein, und zwar sowohl der strategischen als auch der nicht-strategischen. Die Grenzen der Reduktion sind abhängig von der Relevanz der Information (ein Versuch der Operationalisierung wird in Abschnitt 4 beschrieben).

- In vielen Fällen können im Transfer **ambiguitätserhaltende** oder gar **ambiguitätssteigernde Übertragungen** vorgesehen werden. Die Grenzen der Ambiguitätssteigerung richten sich ebenfalls nach der Relevanz der Information.

- Das Prinzip der **Turn-übergreifenden Kohärenz** sollte berücksichtigt werden. Mit Hilfe eines geeigneten Dialoggedächtnisses sollte es möglich sein, die entsprechenden Teile aus Vorgänger-Turns zu finden und zur Entscheidung über die adäquate Wiedergabe heranzuziehen.

- Wie weit die von den DolmetscherInnen verwendeten **Standardannahmen** über stereotype Dialogverläufe und übliche Handlungsweisen der GeschäftspartnerInnen genutzt werden können, scheint eine offene Frage zu sein (s. Abschnitt 2.4). Jedenfalls sollte die bereits vorgesehene Kalendermodellierung hier eine Unterstützung bieten können.

- Ein Konzept einer **Variablen Analysetiefe** könnte für VERBMOBIL nutzbar gemacht werden, um zu entscheiden, welche Problemlösungs-Strategie ggf. anzuwenden ist.[10] Sicher ist, daß nicht in jedem Zweifelsfall ein Klärungsdialog angestoßen werden kann.

3 Kontrolle und Bewertung

3.1 Zielsetzung und Vorgehensweise

Das Arbeitspaket 13.8 ‘Bewertung und Kontrolle maschineller Dolmetschleistungen durch den Nutzer’ untersucht die Kriterien, die von (potentiellen) NutzerInnen an die Leistungen von VERBMOBIL angelegt werden, und erarbeitet mögliche Methoden der Evaluation von VERBMOBIL durch NutzerInnen sowie Modelle zur Überwachung und Beeinflussung dieser Leistungen durch ebendiese NutzerInnen.

Die Orientierung an vorhandenen Erkenntnissen steht bei unserem oben beschriebenen Verfahren der iterativen Hypothesengenerierung an erster Stelle. Da es Erfahrungen mit der systematischen Evaluation maschineller Dolmetschsysteme sowie der speziellen Situation der Mensch-Maschine-Mensch-Interaktion noch nicht gibt, mußten wir hier versuchen, in anderen Bereichen Methoden zu finden, deren Tauglichkeit für die Adaptation an Zwecke der Evaluation und Kontrolle von VERBMOBIL mit einiger Wahrscheinlichkeit angenommen und dann überprüft werden konnte.

Als Ausgangspunkt für unsere Arbeit im Bereich ‘Bewertung’ boten sich vor allem vorhandene Untersuchungen zur Evaluation von maschinellen *Übersetzungs*systemen (MT-Systemen) an. Wichtige Punkte daraus werden im folgenden vorgestellt. Des weiteren gibt es in der Translationswissenschaft eine Tradition der Übersetzungs- bzw. Dolmetschkritik, d.h. der Bewertung menschlicher Translationen.

Auf der Basis derart gewonnener Hypothesen wurden verschiedene Vorstudien konzipiert, die in eine Kriterienaufstellung mündeten. Diese wurde wiederum experimentell überprüft und modifiziert. Dabei wurde zum Teil mit den PartnerInnen an der Universität Hamburg, Institut für Soziologie, zusammengearbeitet.

[10] Genaueres dazu s. [VM-Memo80].

Die einzelnen Versuche und deren Ergebnisse werden im folgenden beschrieben und eine Methodenkritik durchgeführt.

3.2 Erkenntnisse aus der MT-Evaluation

In den Berichten aus dem Bereich der MT-Evaluation wird immer wieder eine Dichotomie zwischen "Nutzer-Evaluation" und "Entwickler-Evaluation" bzw. eine ähnlich gelagerte zwischen Makro- und Mikro-Evaluation (s. [VM-Memo13]) genannt. Eine solche spiegelt sich auch in den Erkenntnissen wider, die bei der Entwicklung von Bewertungskriterien für maschinelle Dolmetschsysteme und deren praktischer Anwendung auf VERBMOBIL erarbeitet wurden. Im folgenden sollen die Überlegungen, die zur Vorgehensweise beim Evaluieren eines maschinellen Dolmetschsystems wie VERBMOBIL aus NutzerInnensicht geführt haben, nochmals zusammenfassend dargestellt werden.
Bei der sog. Entwickler-Evaluation oder Mikro-Evaluation (genaueres zu den Grundlagen in [VM-Memo13] und [VM-Memo26]) wird der übersetzte Text sozusagen als isoliertes Präparat angesehen, das als solches oder im Vergleich zum Originaltext einer Bewertung unterzogen wird. Je nach Evaluationsinteresse kann es sich bei den evaluierten Produkten auch um Teilergebnisse aus einzelnen Moduln handeln. Hierbei kann durch mehrmalige Überprüfung anhand derselben Daten nach Implementierung eines weiteren, neuen Moduls oder Updates besonderes Augenmerk auf die hierdurch erzielte Verbesserung der Systemleistung gerichtet werden.
Für die sog. Nutzer-Evaluation oder auch Makro-Evaluation, das für die Aufgabenstellung von AP 13.8 zu bevorzugende Vorbild, wird gefordert, daß die Qualität des erzeugten Translats, nicht die einzige evaluierte Systemeigenschaft bleiben darf. Evaluiert wird immer das gesamte System in der jeweils gegebenen (Arbeits-) Situation, unter Berücksichtigung des restlichen Arbeitsablaufs. Wichtig sind daher auch Faktoren, die aus der Software-Evaluation im Allgemeinen bekannt sind, die nach ergonomischen Gesichtspunkten vorgeht. Beispiele hierfür sind die Kriterien Transparenz, Verläßlichkeit und Robustheit, die in [VM-Report54] eingehend beschrieben wurden. Sehr ernst zu nehmen ist gerade deshalb die Forderung nach Evaluation unter möglichst realitätsnahen Einsatzbedingungen der zu bewertenden Software. Insbesondere folgenden Faktoren ist Aufmerksamkeit zu schenken:

- Bewertungspersonen müssen (potentielle) AnwenderInnen der entsprechenden Software sein (die Person des 'Evaluators' wurde diskutiert in [VM-Memo26])

- Die Bewertungssituation muß möglichst genau der (zukünftigen) Nutzungssituation unter realistischen Arbeitsbedingungen der NutzerInnen entsprechen (bei MT-Systemen ist das beispielsweise der typische Ablauf eines normalen Übersetzungsauftrags, dessen Verwaltung innerhalb des Büroalltags etc.)

- Es sind ganze, zusammenhängende Texte als Input für die Evaluation zu verwenden. Sie sollen möglichst genau den durch (potentielle) NutzerInnen verwendeten Textsorten entsprechen. .

Der Grund für die zunehmende Bevorzugung der Nutzer-Evaluation liegt in der inhärenten Problematik der fehlenden AnwenderInnen-Perspektive bei der isolierten Bewertung von Outputs durch EntwicklerInnen oder andere Nicht-NutzerInnen. Es ist lt. Erkenntnissen der Translationstheorie davon auszugehen, daß die Qualität einer Translation überhaupt nur im Zusammenhang mit dem sog. Translationsauftrag und den daraus abgeleiteten Translationszielen bewertet werden kann; (zum Begriff des Translationsziels s.a. Abschnitt 4). Dies bedeutet u.a., daß beim Szenario des Gesprächsdolmetschens vor allem die PrimärinteraktantInnen entscheiden können, welches für ihre Zwecke eine adäquate Translation ist.

3.3 Ergebnisse aus explorativen Versuchen an VERBMOBIL-Daten

Von den verschiedenen Evaluationsmethoden, deren Anwendbarkeit auf maschinell gedolmetschte Dialoge im VERBMOBIL-Szenario explorativ getestet wurden, ist keine hundertprozentig kompatibel mit den unter Abschnitt 3.2 geforderten Bedingungen für eine nutzerbezogene Systemevaluation. Ein Grund hierfür liegt darin, daß noch kein System in dem dort beschriebenen Sinne existiert, d.h. daß VERBMOBIL noch nicht als fertige Einheit vorliegt. Ein weiterer, damit zusammenhängender Grund ist, daß auch keine "NutzerInnen" in dem o.g. Sinne existieren, sondern allenfalls potentielle NutzerInnen. Die Zusammensetzung dieser Gruppe basiert aus naheliegenden Gründen auf Hypothesen, und so müssen auch Untersuchungsergebnisse zu deren Verhalten im Umgang mit VERBMOBIL und deren Einstellungen zu dieser Technologie entsprechend vorläufig bleiben.
Die Defizite der einzelnen Ansätze wurden jeweils in den entsprechenden Einzelpublikationen diskutiert (s.u.). Es wurde jedoch versucht, durch Annäherung nach dem Prinzip der Triangulation einzelne Aspekte isoliert zu realisieren und daraus Rückschlüsse auf Evaluationen unter Idealbedingungen zu ziehen.

3.3.1 Bewertung maschineller Übersetzungen

In diesem Vorversuch, der der Kriterienfindung und -priorisierung diente, wurden durch studentische Versuchspersonen maschinell übersetzte Sätze bewertet. Die Bewertung diente lediglich als Basis für eine folgende Diskussion über zugrundeliegende Kriterien bei Bewertung einer maschinellen Translation im allgemeinen.

Das Ergebnis war eine vorläufige Kriterienliste. Zu Details siehe [VM-Report54, Abschnitt 3.2] und [VM-Memo27, Abschnitte 2 und 3].

3.3.2 Bewertung menschlicher Verdolmetschungen

Bei diesem Versuch handelte es sich um die Bewertung der *menschlichen* Verdolmetschung eines TP13-Gesprächs[11] anhand einer Videoaufzeichnung und des Transkripts. Dieser Versuchsaufbau war szenario-näher als der im vorigen Abschnitt beschriebene; dadurch sollten sich Unterschiede zu den bei Bewertung von maschineller *Übersetzung* angelegten Kriterien herausstellen. Nachteil dieser Methode war jedoch die Gefahr der Beeinflussung der Versuchspersonen (Sekundär-VPn) durch das Verhalten der Personen (Primär-VPn) im Video. Auch hier wurden die Bewertungen der Versuchspersonen im Plenum diskutiert.
Das Ergebnis war eine verfeinerte, erweiterte und veränderte Kriterienliste (s. Abbildung 2). Zu Details siehe [VM-Report54, Abschnitt 3.3] und [VM-Memo27, Abschnitte 2 und 3].

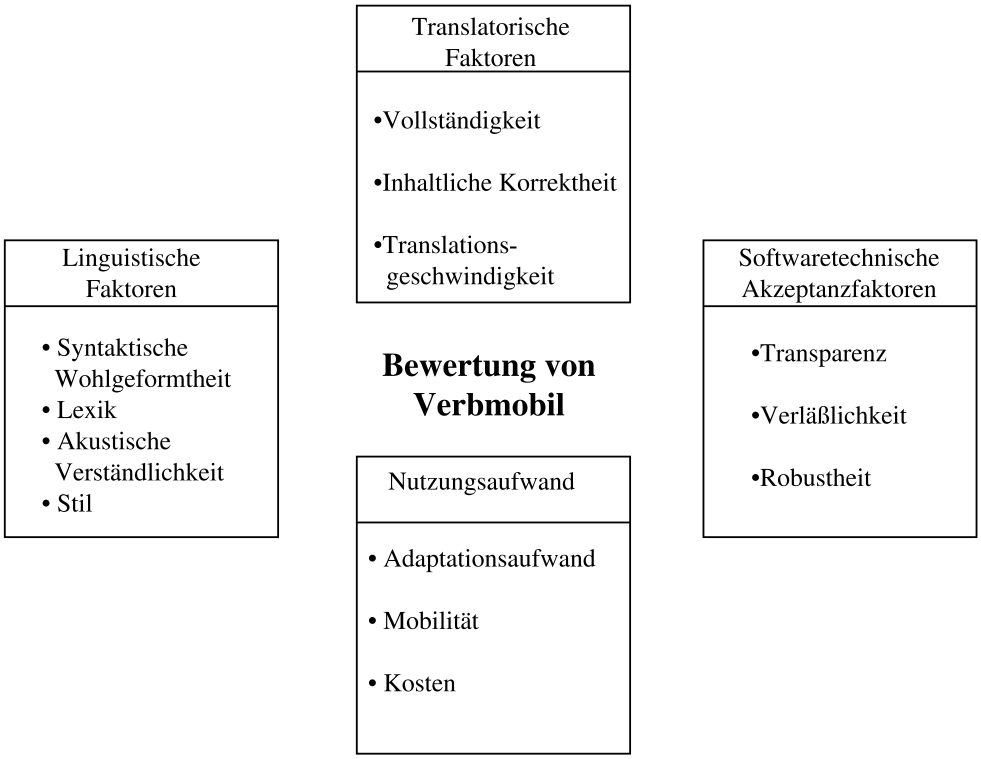

Abbildung 2: Erste Version der Kriterienliste für die Bewertung von VERBMOBIL durch den/die NutzerIn

[11] TP13-Gespräch 12, s. [VM-Memo18].

3.3.3 Auswertung der Erfahrungen mit dem Demonstrator

Entsprechend dem Arbeitsplan hatten wir unsere Ergebnisse am VERBMOBIL-Demonstrator zu überprüfen. Diese Überprüfung ergab nicht viel Neues, da ein Großteil der für ein fertiggestelltes VERBMOBIL-System erarbeiteten Kriterien nicht für den Demonstrator in frage kam und zuviele Nicht-Translationen erfolgten (Näheres s. [VM-Memo80]).
Auch die gesammelten Daten (Demonstrator-Translate) konnten aus diesen Gründen nicht für reale Studien weiterverwendet werden.

3.3.4 Befragung von Versuchspersonen bei Simulator-Experimenten

Bei den Experimenten handelte es sich um die durch MitarbeiterInnen des Instituts für Soziologie, Universität Hamburg (AP 13.5, 13.7, 13.9), durchgeführte Hauptstudie (s. [VM-Report94]).
Mitarbeiterinnen des Arbeitspakets 13.8 waren an der Durchführung dieser Studie beteiligt, denn der Versuchsansatz der Simulation (verschiedentlich auch 'Wizard-of-Oz-Szenario' genannt), erfüllte fast ideal die unter Abschnitt 3.2 beschriebene Forderung nach Identität der Bewertungspersonen mit NutzerInnen des untersuchten Systems. Eine tatsächliche Nutzung eines von den Versuchspersonen (VPn) als "VERBMOBIL" anerkannten Systems fand statt, und dieselben Personen konnten zu Kriterien befragt werden. Es konnten allerdings keine einzelnen Turns bewertet werden, es wurden nur Prioritäten zwischen Bewertungskriterien diskutiert.
Das Ergebnis war eine Priorisierung zwischen einigen besonders wichtigen Kriterien (s.a. [VM-Memo98, S.15ff.] und [VM-Memo88]) in der folgenden Reihenfolge:

1. Geschwindigkeit
2. Korrektheit
3. Vollständigkeit
4. Stil

Weitere interessante Erkenntnisse bezogen sich auf die Tendenz, nur einige wenige, umfassender definierte Kriterien (also z.B. "sprachliche Fehlerlosigkeit" anstatt "Lexik" oder auch "Wortschatz" und "Syntax" oder "Grammatik") zuzulassen, sowie die unter NutzerInnen vorherrschenden Vorstellungen z.B. zu den Begriffen "Vollständigkeit" und "Korrektheit". Erkenntnisse aus dieser Studie haben wesentlich dazu beigetragen, daß die Kriterienliste praxisnäher gestaltet werden konnte. Die Ergebnisse flossen in die verbesserte Kriterienaufstellung ein, die in Abschnitt 3.6.1 diskutiert wird.

3.3.5 Hauptuntersuchung: Gruppendiskussion anhand Videoaufzeichnung einer Simulation

In dieser Haupt-Untersuchung sollten die Ergebnisse aus den Vorversuchen in einer möglichst szenario-nahen Situation mit echten potentiellen NutzerInnen, d.h. Personen aus Industrie und Wirtschaft, die Dolmetschbedarf haben, verifiziert werden (zur Gewinnung solcher VPn s. a. ([VM-Report94, S. 2f.]). Es hatte sich inzwischen erwiesen, daß ein solcher Personenkreis extrem schwer zu motivieren ist, unter den gegebenen Bedingungen an umfangreichen Studien teilzunehmen. Näheres zu dieser Untersuchung siehe [VM-Memo98].

Das Ergebnis war eine zweite Version der Liste von Bewertungskriterien. Es ergab sich v.a. eine Neudefinition einiger Kriterien und die stärkere Annäherung an intuitive Vorstellungen translationsunerfahrener NutzerInnen. Die Notwendigkeit einer gröberen Granularität der Kriterien wurde bestätigt. Außerdem wurden interessante Nutzermeinungen zu Leistungsmerkmalen eines maschinellen Dolmetschsystems elizitiert.

Eine sich bereits andeutende Tendenz zur zweistufigen Priorisierung von Kriterien unter NutzerInnen hat sich in dieser Untersuchung noch bestätigt. Siehe hierzu Abschnitt 3.6.1.

Die sich hieraus unter Berücksichtigung der soziologischen Ergebnisse aus Abschnitt 3.3.4 ergebende endgültige Kriterienaufstellung wird ebenfalls unter Abschnitt 3.6.1 beschrieben.

3.3.6 Bewertungsversuch zum Kriterium der Vollständigkeit

In einem weiterführenden Versuch sollten einige der erarbeiteten Methoden-Details angewendet werden. In den vorhergehenden Versuchen handelte es sich nur darum, Kriterien herauszuarbeiten und zu priorisieren. Nun sollte das als relevant erkannte Kriterium der Vollständigkeit anhand von präparierten Translationen in einem Rating-Experiment angewandt und die Einstellung der VPn zu reduzierten Translaten untersucht werden.

Entsprechend den Methodik-Erkenntnissen aus früheren Versuchen beschränkten wir uns auf eine Einzelbefragung und auf ein einfaches Rating, in dem zwei schriftlich vorliegende Translationen desselben Turns miteinander verglichen werden sollten. Die VPn sollten entscheiden, welches der beiden Translate besser sei oder ob beide gleich gut seien. Das Kriterium der Vollständigkeit wurde hierbei nicht expliziert, das Ausgangsmaterial[12] wurde jedoch so manipuliert, daß jeweils Paarungen von reduzierten und nicht reduzierten Translaten präsentiert wurden. Bei der Versuchsvorbereitung wurde die Aufmerksamkeit der Versuchspersonen gezielt auf den Sachverhalt 'wörtliche' vs. 'freie' Translation gelenkt. Die Gründe für die Entscheidung sollten zu jedem Turn angegeben werden. In ei-

[12]Simulator-Gespräch 21 wie in [VM-Memo24] beschrieben.

nem abschließenden Interview wurde außerdem abgefragt, welche Art von Translat *generell* als besser angesehen wird.
Vertiefend wurde anhand zweier bereits bewerteter Turns eruiert, welche Abstufung von Reduktion (drei weitere Translatvarianten) der persönlichen Bewertungsintuition der VPn am nächsten kommt. Diese Varianten unterschieden sich darin, daß sie mehr oder weniger der in der Originaläußerung realisierten bzw. im jeweiligen Translationsziel (s. Abschnitt 4) geforderten Dialogakte enthielten. Die Auswertung dieses Teilversuchs hat nährere Einsichten in die Nutzbarkeit des Konzepts des Translationsziels für die NutzerInnenbewertung von VERBMOBIL ergeben.
Bei den Versuchspersonen handelte es sich um 20 wie unter Abschnitt 3.3.5 beschriebene potentielle NutzerInnen. Vor der schriftlichen Bewertung auf standardisierten Bewertungsbögen gab es eine akustische Präsentation (Tonaufnahme) von Ausschnitten des Dialogs, damit die VPn sich den Ablauf eines maschinegemittelten Dialogs vorstellen konnten.
Ergebnisse der Gesamtbewertung waren folgende:

- **Ausgeprägte individuelle Variation im Akzeptanzverhalten von Reduktion.** In der *generellen* Schlußbewertung waren zwei relativ gleich große Gruppen von VPn zu erkennen: Entweder wurde *generell* die ausführlichere Variante als besser eingeschätzt (9 Personen) oder die reduzierte (8 Personen). Nur eine Person hätte lieber einen Mittelweg gehabt, zwei Personen wollten sich dagegen die Entscheidung je nach Situation vorbehalten.

  Als Gründe wurden auf beiden Seiten ähnliche Argumente vorgebracht: Die eine Variante sei ausführlicher, die andere gebe nur 'das Nötigste' wieder. Je nach Auffassung wurde das eine oder das andere als in geschäftlichen Terminabsprachen notwendig erachtet.

  Die Entscheidung für die eine oder andere Variante scheint ganz stark von persönlichen Vorlieben abzuhängen. Bei den Personen, die Wahlmöglichkeit fordern, spielt es bei der Entscheidung eine große Rolle, als wie offiziell die Gesprächssituation eingeschätzt wird.

- **Unterschiedliche Grundeinstellungen gegenüber der Technologie.** Generelles Vertrauen oder generelles Mißtrauen gegenüber einer "solchen Maschine" war ebenfalls von großer Bedeutung für die Entscheidung für oder gegen die reduzierte Variante, wobei einige wenige der - translationsunerfahrenen - VPn zu glauben schienen, daß eine Wort-zu-Wort-Verdolmetschung von Dialogen praktikabel ist.

- **Das Dialogsprachen-Modell spielt eine Rolle.** Bei den Befürwortern der reduzierten Version gaben einige ihr eingeschränktes Verständnis der Dialogsprache Englisch als Grund an.

Die Angaben der VPn bestätigten die schon in der Hauptuntersuchung von der dort befragten ExpertInnengruppe (s. Abschnitt 3.3.5) gemachten Aussagen.

3.4 Fallbeispiel: End-to-End-Evaluation des Forschungsprototypen

Im Vorfeld der Abnahme des Forschungsprotoypen am Ende der ersten Projektphase fand eine Großevaluation der Gesamtleistung von VERBMOBIL statt, an der u.a. Mitarbeiterinnen des AP 13.8 als Bewertungspersonen teilnahmen.
Es handelte sich hierbei um eine binäre Bewertung. Der Input, einzelne transkribierte Turns (z.T. in Reihenfolge und Zusammenhang von Gesamtdialogen), wurde in einem online-Formular dem jeweiligen Output des Forschungsprototypen gegenübergestellt. Die Bewertungspersonen hatten zu entscheiden, ob die Translation adäquat ist oder nicht. Die Ergebnisse wurden gespeichert und von der Systemgruppe statistisch ausgewertet.
Auf der Basis von groben Absprachen zwischen den Bewertungspersonen splittete sich die Bewertung 'adäquat' in zwei Kriterien auf, und zwar 'inhaltlich korrekt' und 'vollständig'.
Diese Evaluation durch TranslatorInnen anhand von Transkripten war u.E. eine reine Entwickler-Evaluation, da sie keine einzige der o.g. Forderungen an eine Nutzer-Evaluation erfüllen konnte. Die Methodik war alleine von der verfügbaren Form ausreichender Daten und anderen technischen Gegebenheiten diktiert sowie von einem erheblichen Zeitdruck bestimmt: In der zur Verfügung stehenden Zeit hätten Evaluationsergebnisse mit 'echten' NutzerInnen in authentischen Situationen nicht in ausreichender Menge erhoben und vor allem nicht quantitativ ausgewertet werden können.
Es sollte zwar versucht werden, 'Fachleute' zur Bewertung der Adäquatheit der Translationen heranzuziehen, die nicht aus dem engeren EntwicklerInnenkreis stammen, aber es handelte sich hierbei nicht um (potentielle) NutzerInnen. Lediglich die grobe Rasterung der Kriterien (adäquat - inadäquat) stellte eine Konzession an die NutzerInnenperspektive dar. Der Sinn der Evaluation bestand vielmehr darin, sich von den bislang erfolgten einzelmodul-orientierten Teilevaluationen zu entfernen und vor allem die Gelegenheit zu haben, ins Gesamtsystem eingebettete, verschiedene Konfigurationen zu testen sowie die Ergebnisse 'flacher' und 'tiefer' Analyse- und Transferverfahren miteinander zu vergleichen. Außerdem diente diese Evaluation dazu, die Erfüllung eines selbstgesetzten, quantitativ formulierten Entwicklerziels (bestimmter Prozentsatz an adäquaten Translaten) sowie den Fortschritt dorthin zu überprüfen.
Es muß betont werden, daß gerade die Quantifizierung nach Prozentsätzen korrekter Translate nicht nur für die Evaluation maschineller Dolmetschsysteme als problematisch anzusehen ist; sie ist auch aus der MT-Evaluation als problematisch bekannt. Schon ohne die der Spontansprache inhärenten Segmentierungs- und

Kategorisierungsprobleme (Prozentsatz wovon? Wörter? Sätze? Turns? "Segmente"? Dialogakte?) ist bei einem solchen quantitativen Ansatz, der zwangsläufig fehlerbasiert ist, der zugrundeliegende Fehlerbegriff (Wie schwer ist der Fehler? In welchem Modul hat er seine Ursache?) für den/die interessierte EntwicklerIn ein Anlaß dazu, in solche einfache, dichotome Kriterienpaare wieder Abstufungen bzw. Relativierungen einzuführen, die zusätzlich Auswertungsarbeit erfordern und neue Bewertungsunsicherheit einbringen.[13]
Als eine weitere Fehlerquelle stellte sich die mangelnde Vorbereitung der Bewertungspersonen heraus. Anhand von Stichproben war festzustellen, daß die individuellen Auffassungen von "adäquat" bei der gegebenen groben Zweiteilung in Kriterien trotz Absprache divergierten. Bei weiteren End-to-End-Evaluationen ist deshalb die Zugrundelegung von Translationszielen (s. Abschnitt 4) sowie eine vorhergehende Schulung oder kooperative Testevaluation mit allen Bewertungspersonen zu empfehlen.

3.5 NutzerInnen-Kontrolle von VERBMOBIL

Ein zweiter Arbeitsbereich von AP 13.8 bezog sich auf die Kontrolle von VERBMOBIL durch die NutzerInnen. In diesem Zusammenhang wurde ein Kontrollmodell vorgestellt (s. [VM-Memo13]), das sich auf Erkenntnisse aus der MT-Evaluation stützt. Es sieht akustisch-visuell-taktile Mensch-Maschine-Interaktion gegenüber rein akustischer Interaktion vor. Dieses Konzept fordert NutzerInnenkontrolle der Translation des eigenen Gesprächsbeitrags *vor* der akustischen Synthese und damit vor der Wahrnehmung des jeweils anderen Gesprächspartners, und ist seitdem durch Anwendung in anderen maschinellen Dolmetschsystemen wie "JANUS" in seiner Relevanz bestätigt worden. Variationen dieses Konzepts sind seither auf zahlreichen Projektworkshops diskutiert worden, beispielsweise als ausgangssprachige, visuelle Wiedergabe des Erkenner- oder Analyseergebnisses zur Verifikation durch den/die NutzerIn oder als zielsprachige Präsentation von Transfer-/Generierungsvarianten zur Auswahl vor der Synthese.[14]
Ein großer Vorteil dieses Modells besteht u.E. darin, daß es die Verschiedenheit der maschinell gemittelten gegenüber der menschlich gemittelten mehrsprachigen

[13] Beispiele beim VERBMOBIL-Forschungsprototypen: Wenn durch einen systematischen Fehler einzelne Aussagesätze als Fragesätze, ansonsten aber korrekt wiedergegeben werden, kann es sich, wenn Wissen über die Behebbarkeit dieses Fehlers vorliegt, für die EvaluatorIn um einen minderen Fehler handeln. Wenn bei einer Nicht-Translation der Grund für das Scheitern einzelnen Moduln zugeordnet werden kann ('no syntax', 'no gener'...), kann das eine für die EntwicklerIn interessante Abstufung sein. Es muß jeweils genau überlegt werden, ob es im jeweiligen Evaluationsinteresse ist, diese Abstufungen zu erhalten oder aber die einfache Dichotomie beizubehalten. Im zweiten Falle ist es wichtig, daß die jeweiligen Bewertungspersonen sich von diesen Zusatzinformationen in ihrer Bewertung nicht beeinflussen lassen.

[14] Das setzt allerdings ein Dialogsprachenmodell voraus, das für Phase II vn VERBMOBIL in Frage gestellt ist.

Dialogsituation hervorhebt und so Transparenz und Akzeptanz fördert.
Aus den Untersuchungen des Arbeitspakets 13.3 zum Aspekt der Hypothesenrevision läßt sich für die NutzerInnenkontrolle eines VERBMOBIL-Systems außerdem schließen, daß den NutzerInnen nicht zuviel an Korrekturvorgängen zugemutet werden darf, weil dies in der ohnehin komplexen Kommunikationssituation eine zu hohe zusätzliche Belastung bedeuten würde.

3.6 Schlußfolgerungen

3.6.1 Kriterienaufstellung:

Aufgrund der weiter oben geschilderten Gegebenheiten haben wir uns bei den weiterführenden Untersuchungen auf einige der in den Voruntersuchungen aufgestellten Kriterien konzentriert. Dadurch konnten vor allem diejenigen Kriterien näher untersucht werden, die aus den meisten Befragungen mit der höchsten Priorität hervorgingen. Die verbleibenden Kriterien aus der ersten Liste (s. Abbildung 2) werden dadurch nicht obsolet, sondern sollten ebenfalls näher untersucht werden. Insbesondere der Einfluß der akustischen Verständlichkeit auf die Wahrnehmung der höher priorisierten Kriterien könnte interessant sein. Außerdem werden bei Vorliegen eines fertigen VERBMOBIL-Systems die Kriterien der Gruppe "Nutzungsaufwand" relevant.
Die weiterführenden Untersuchungen ergaben, anstatt der erwarteten Prioritätenliste von Kriterien, eine zweistufig priorisierte Aufstellung verschiedener Systemeigenschaften.
Wir verstehen unter einer zweistufigen Priorisierung, daß die meisten (translationsunerfahrenen) potentiellen NutzerInnen die möglichen Bewertungskriterien für ein maschinelles Dolmetschsystem nicht in einer Reihe untereinanderstellen, sondern einige als grundlegend und andere als optional ansehen.
Die Erfüllung der **grundlegenden Kriterien** steht außerhalb jeglicher Diskussion. Werden diese Kritierien nicht erfüllt, ist das System nicht akzeptiert und gilt schlichtweg nicht als Dolmetschsystem.[15] Es handelt sich somit nicht um Kriterien im eigentlichen Sinne; sie werden im folgenden als "Voraussetzungen" bezeichnet. Zu diesen Voraussetzungen gehören eine inhaltlich **korrekte** Übersetzung sowie die softwaretechnischen Akzeptanzfaktoren nach Abbildung 2, dort v.a. die Robustheit. Es gibt Anhaltspunkte dafür, daß die VPn (ähnlich wie TranslatorInnen bei der Selbstbewertung), für die Beurteilung, welche Translate als "korrekt" gelten können (s. Abschnitt 4), eine Vorstellung im Kopf haben, die unserem Translationsziel nahekommt: Es müssen die relevanten Dialogakttypen und propositionalen Bestandteile mit hinreichender Präzision wiedergegeben

[15] Die Kriterien stehen somit nicht für eine Nuetzerevaluation zur Verfügung, sind jedoch auf der Ebene der Entwickler- bzw. Mikroevaluation weiter relevant.

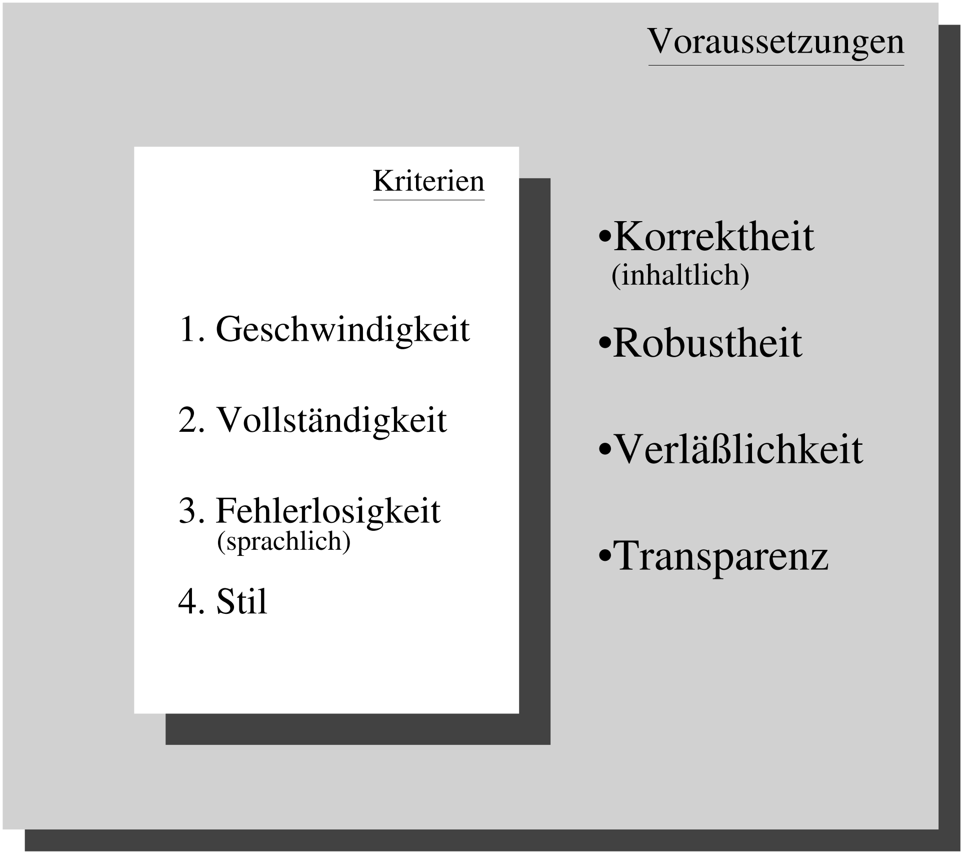

Abbildung 3: Kriterienaufstellung für die Bewertung von VERBMOBIL durch den/die NutzerIn

werden. Es geht also um die Erfüllung der qualitativen Anforderungen des Translationsziels.
Die Erfüllung oder Nichterfüllung der **"optionalen"** Kriterien macht dagegen den Unterschied zwischen einem guten und einem schlechten System aus. Nur in dieser Liste ergibt eine Priorisierung einen Sinn. Es handelt sich um folgende Kriterien:

Translationsgeschwindigkeit: Die Zeit, die zwischen dem Ende der Eingabe und dem Beginn der Ausgabe verstreicht.

Vollständigkeit: Ob eine Translation "vollständig" ist, kann daran gemessen werden, welche Elemente des ausgangssprachlichen Segments im zielsprachlichen Segment realisiert werden. Als Mindestforderung kann hierbei angesehen werden, daß sämtliche im Translationsziel (s. Abschnitt 4) enthalten Dialogakte und propositionalen Bestandteile wiedergegeben sein müssen. Es geht hier also um die Erfüllung der quantitativen Anforderungen des Translationsziels.

Fehlerlosigkeit: Syntaktische Wohlgeformtheit und adäquate Lexik.

Stil: Hierunter werden meistens Elemente der Höflichkeit wie z.B. Abtönungselemente verstanden.

Aus den abschließenden Untersuchungen zur Vollständigkeit geht hervor, daß es eine ausgesprochene Fraktionierung der Versuchspersonen, und vermutlich der potentiellen NutzerInnen insgesamt, in zwei Gruppen gibt: Grundsätzliche Forderung nach oder grundsätzliche Ablehnung von reduzierenden Verdolmetschungen (dies betrifft *nicht* die Reduktion von Häsitationsphänomenen, denn deren Notwendigkeit steht außer Frage). Die Definition des Kriteriums "Vollständigkeit" wird also subjektiv unterschiedlich vorgenommen. Für die VertreterInnen einer eher sinngemäßen Verdolmetschung liegt die Meßlatte einer vollständigen Verdolmetschung in etwa bei der Menge der im Translationsziel definierten Elemente. VertreterInnen einer eher wörtlichen Verdolmetschung möchten mehr Elemente des ausgangssprachlichen Segments realisiert wissen, als im Translationsziel definiert sind.
Es ist dennoch sinnvoll, sich bei der Definition eines für VERBMOBIL operationalisierbaren Kriteriums "Vollständigkeit" am Konzept des Translationsziels zu orientieren. Dies gilt schon deshalb, weil es schwierig ist, eine vollständige Verdolmetschung im Sinne von "eher wörtlich" systematisch von einer redundanten Verdolmetschung abzugrenzen, die nicht mehr akzeptabel ist, weil sie Spuren des ursprünglichen Sprachproduktionsprozesses übernimmt. Ein zweiter Grund ist, daß eine Über-Erfüllung des Vollständigkeitskriteriums erfahrungsgemäß zu Lasten anderer Kriterien bzw. Voraussetzungen geht.

3.6.2 Methodenkritik:

Folgende Probleme erschweren die Entwicklung einer adäquaten Makro-Evaluationsmethode für VERBMOBIL auf seinem Entwicklungsstand vom Ende der Phase 1:

- Bei den Inputdaten ergibt sich folgendes Dilemma: In einer authentischen Situation von potentiellen NutzerInnen gesprochene Dialoge oder auch Einzeläußerungen bewegen sich lexikalisch zu oft außerhalb der Wortliste des VERBMOBIL-Forschungsprototypen ("out of vocabulary"), das heißt, sie enthalten Lexik, die bis jetzt nicht trainiert wurde und nicht erkannt werden kann. Solche Äußerungen können zur Evaluation nicht herangezogen werden, da eine Messung der Leistungsfähigkeit durch unbekannte Wörter verzerrt wird und es nicht die Möglichkeit gibt, wie z.B. bei einem MT-System, dieses Defizit während der Evaluation kurzfristig zu kompensieren.

- Eine Lösungsmöglichkeit wäre, nachträglich aus einer entsprechend größeren Menge Datenmaterials diejenigen Turns bzw. Segmente auszufiltern, die der Wortliste entsprechen. Nur diese würden dann einer Evaluation zugeführt. Neben der Tatsache, daß diese Methode sehr materialaufwendig ist, ergibt sich dann das Problem, daß man niemals zusammenhängende Dialoge oder auch nur Dialogteile zur Bewertung hätte. Die o.g. Forderung nach Evaluation anhand zusammenhängender Texte könnte nicht erfüllt werden. Turn-übergreifende Leistungsmerkmale bestimmter Systemmoduln (z.B. des Dialoggedächtnisses) könnten nicht realistisch zum Bewertungsergebnis beitragen. Auch hiermit würden also Bewertungsergebnisse verfälscht.

- Läßt man jedoch Dialoge bzw. Turns sprechen oder ablesen, die ausschließlich "erlaubte" Lexik enthalten, handelt es sich zwangsläufig nicht mehr um Spontansprache. Da diese jedoch unverzichtbarer Bestandteil des VERBMOBIL-Szenarios ist, bewegen sich auch solcherart aufgenommene Evaluationsdaten "out of scenario", d.h. sie sind realitätsfern.

Es wird sofort klar, daß es sich bei Makro-Evaluationsmethoden, die den o.g. Forderungen genügen, um Ideale handelt, die zumindest bei noch in Entwicklung befindlichen Maschinellen Dolmetschsystemen - wie VERBMOBIL zum jetzigen Zeitpunkt - nicht erfüllbar sind. Es müssen also Kompromisse gefunden werden, d.h. Methoden, die sich von verschiedenen Seiten dem Ideal annähern und in ihrer Gesamtheit oder zeitlichen Abfolge ein hinreichend aufschlußreiches Bild von der Systemleistung geben. Auch Methoden der reinen Mikro-Evaluation bzw. Entwickler-Evaluation gehören sicherlich aufgrund des frühen Entwicklungsstandes von VERBMOBIL in einem bestimmten Umfang dazu. Dies entspricht dem

Prinzip der Triangulation, dessen Anwendung sich für unsere Zwecke als brauchbar erwiesen hat. Dieses Prinzip bedeutet aber auch, daß jede der erarbeiteten Methoden zur experimentellen Überprüfung von Kriterienhypothesen nur in einem begrenzten Umfang verwertbare Ergebnisse erbringen kann. Es soll hier eine kurze Bewertung der Anwendbarkeit jedes verwendeten Prinzips erfolgen.

Bewertung maschineller Übersetzungen bzw. verschriftlichter (maschineller) Verdolmetschungen

Siehe Abschnitt 3.3.1, 3.4 und 3.3.6.

Vorteile: Das Ausgangsmaterial ist leicht und schnell variierbar. Vergleichbare Bewertungen sind leicht möglich. VPn werden zeitlich wenig belastet; dadurch größere Anzahl an VPn möglich. Statistisch relevante Datenmengen sind schneller erreichbar. Laufende bzw. wiederholte Bewertung ist praktikabel.

Nachteile: Akustische Qualitäten des Datenmaterials sind schwer wiedergebbar. Das Datenmaterial ist nicht szenario-nah. Der (Echt-) Zeitbezug ist nicht gegeben. Der Gesamtzusammenhang des zu bewertenden Textes geht leicht verloren; statt dessen werden Einzelsätze bewertet. Der Blick auf die Dialogsituation ist versperrt.

Bewertung menschlicher Gesprächsverdolmetschungen durch Sekundär-VPn

Siehe Abschnitt 3.3.2.

Vorteile: Durch die Befragung von Sekundär-Versuchspersonen mit Hilfe von Aufzeichnungen (im Gegensatz zu den in der Aufzeichnung sicht- oder hörbaren Primär-Versuchspersonen, die am Dialog beteiligt waren) kann die gleiche Situation von mehreren Personen bewertet werden. Simulator oder VERBMOBIL-Version braucht nicht vorzuliegen. Gespräche können authentisch spontansprachlich oder zumindest glaubwürdige Rollenspiele sein. Die wenig bekannte Gesprächsdolmetsch-Situation kann den VPn gut nahegebracht werden. Translate können in begrenztem Umfang manipuliert werden.

Nachteile: Video- oder zumindest Audioaufzeichnung ist nötig. Vergleichende Bewertung alternativer Translationsstrategien ist schwerfällig und aufwendig. Manipulationen an der Dolmetschleistung verfälschen die Authentizität der Situation.

Bewertung simulierter maschineller Verdolmetschungen durch Sekundär-VPn

Siehe Abschnitt 3.3.5.

Vorteile: Der Simulator bereitet gegenüber dem "echten" VERBMOBIL keine Vokabular-Probleme. Output-Daten können in begrenztem Umfang manipuliert werden, um bestimmte Gesichtspunkte im Akzeptanzverhalten zu testen.

Die Primär-VPn brauchen nicht zusätzlich befragt zu werden. Es können mehrere Sekundär-VPn mit demselben Reizmaterial konfrontiert werden. Theoretisch dadurch größere VP-Zahl. Sehr nah an der angestrebten Idealmethode. Eignet sich gut zur Kriterienfindung bei einer neuen Technologie.

Nachteile: Technisch aufwendig, da Videoaufzeichnungen nötig. Das beobachtete Verhalten der Primär-VPn beeinflußt die Sekundär-VPn. Die Anzahl der potentiellen VPn ist praktisch doch nicht wesentlich höher, da auch die Sekundär-VPn zeitlich stark belastet werden. Eignet sich weniger zur laufenden Bewertung bestehender Technologien, da Datenmanipulation (z.B. reduziertes vs. ausführliches Translat) und vergleichende Bewertung alternativer Systemleistungen schwerfällig und aufwendig sind.

Es ist bei der Variation verschiedener Evaluationsmethoden sehr wichtig, das Ziel einer (fast) 'idealen' Nutzer-Evaluation nicht aus den Augen zu verlieren und die Realitätsnähe einer solchen Evaluation abwechselnd von unterschiedlichen Positionen aus zu approximieren. Auch eine Evaluation des tatsächlich installierten Systems durch potentielle NutzerInnen ist immer wieder vorzunehmen. Dadurch kann vermieden werden, daß Leistungsmerkmale wie NutzerInnenfreundlichkeit der Oberfläche, Kontrollmöglichkeiten, Ausgestaltung von Klärungsdialogen und v.a. die Entwicklung der Mensch-Maschine-Mensch-Kommunikation unter dem Einfluß dieser neuen Technologie gegenüber der rein translatorischen Leistung vernachlässigt werden.

3.6.3 Zur Kontrolle durch die NutzerInnen

Es ist kein allgemeiner Trend zur Bevorzugung einer bestimmten Kontrollvariante unter den potentiellen NutzerInnen erkennbar. Im Gegenteil: Es scheint genau zwei entgegengesetzte Meinungen bzw. Nutzungskonzepte zu Systemen wie VERBMOBIL zu geben (Äußerungen aus der Hauptuntersuchung, s. 3.3.5):

1. Das System soll vollautomatisch funktionieren, es soll möglichst unauffällig sein und ein Minimum an Bedienungsaufwand erfordern. Man soll es nicht sehen, sondern nur hören.

2. Das System soll mir verschiedene Eingriffsmöglichkeiten bieten. Ich will es nur als Unterstützung benutzen und seine Performanz kontrollieren können.

Diese Zweiteilung scheint gekoppelt zu sein mit der Grundeinstellung dazu, was als 'gutes' Translat angesehen wird (s. Abschnitt 3.3.6): Wenn die VP den Eindruck hat, daß Nutzung des Geräts eine neue Situation ist, man die anderen Gesprächspartner noch nicht kennt, das Gespräch wichtig ist, oder man eher ausführliche Translate haben möchte, möchte man auch mehr Kontrollmöglichkeiten. Wird das 'Team' aus Dialogpartnern und Maschine als gut eingespielt eingeschätzt oder/und bevorzugt man reduzierte Translate, akzeptiert man eine vollautomatische "black box". Dies entspricht entweder einer persönlichen Grundhaltung oder soll mit ein und demselben System variabel einstellbar bleiben.

4 Translationsziele für VERBMOBIL

In den vorangehenden Abschnitten hat sich mehrfach gezeigt, daß eine Art "Meßlatte" für adäquate Verdolmetschungen aus theoretischen und praktischen Gründen unverzichtbar ist. Erstaunlicherweise liefert die Translationswissenschaft kein geeignetes Instrumentarium, das sich auf den hier interessierenden Fall von menschlichen oder maschinellen Dolmetschleistungen anwenden ließe. In der sogenannten Übersetzungskritik werden einerseits mikrostrukturelle Probleme wie fehlende Lexem-Äquivalenzen bzw. lexikalische Lücken in der Zielsprache abgehandelt (die für das maschinelle Dolmetschen natürlich nicht unwichtig sind, aber für dessen Zwecke als Kategorien keinesfalls ausreichen) - andererseits werden makrostrukturelle Gesichtspunkte wie "Textfunktion", "Kommunikationsziel", "Übersetzungsauftrag" diskutiert, die zwar ebenfalls wichtig sind, sich aber nicht operationalisieren lassen. Was fehlt, sind Adäquatheitskriterien auf der Meso-Ebene, die z.B. erlauben würden, die Angemessenheit einer Translation zu einer Äußerung oder einem Redebeitrag im Dialog in einer bestimmten Situation verläßlich zu beurteilen, und die so zwischen Mikro- und Makro-Ebene vermitteln könnten.[16]

Nun konnte aber anhand von umfangreichen Datenanalysen zu menschlichen Verdolmetschungen im VERBMOBIL-Szenario festgestellt werden, daß menschliche DolmetscherInnen hierzu eine recht klare Intuition zu haben scheinen, was in dieser Situation relevant ist und was nicht oder weniger relevant ist. Das Reduktionsbeispiel in Abschnitt 2.3 ist prototypisch dafür.

Auf der Grundlage dieser Beobachtungen entwickelten wir in Zusammenarbeit mit den PartnerInnen an der Universität Hamburg (AP 13.1 'Dolmetschstrategien') und der TU Berlin (TP 11 'Semantische Auswertung') das Konzept des "Translationsziels", das ein Adäquatheitskriterium auf der Meso-Ebene darstellt.

[16] Siehe hierzu die Dissertation von Birte Schmitz (TP 11), TU Berlin.

Die erste Version eines Translationsziels für VERBMOBIL ist folgende:[17]

1. Wiedergabe des Dialogakttyps
2. Präzise Terminbeschreibung
3. Mittlerer Höflichkeitslevel

Dieses Translationsziel bezieht sich auf die Verhandlungsphase einer Terminabsprache.
In der weiteren Arbeit wurde versucht, dieses Konzept zu verallgemeinern, was zu folgenden Schlußfolgerungen führte:

- die Dialogakt-Typen müssen voraussichtlich etwas feiner differenziert werden als es derzeit im Gesamtprojekt angenommen wird (z.B. müssen vermutlich eingeschränkte und absolute Ablehnung/Akzeptanz unterschieden werden);
- die erwarteten Bestandteile des propositionalen Gehalts, die unbedingt mit einem (genauer zu beschreibenden) Präzisionsgrad wiedergegeben werden müssen, hängen offensichtlich vom Dialogakt-Typ ab und können auf dieser Grundlage genauer definiert werden;
- es hängt ebenfalls vom Dialogakt-Typ ab, welche Verbalisierung als Ausprägung eines "mittleren Höflichkeitslevels" angesehen werden kann: bestimmte Dialogakte sind selbst schon als "Höflichkeitselemente" zu betrachten und verlangen dementsprechend keine zusätzlichen höflichen Elaborationen, während andere Dialogakte für den Gesprächsverlauf besonders kritisch sind und daher besondere Anstrengungen in Richtung Höflichkeit verlangen.

Die Operationalisierbarkeit von Translationszielen wird z.B. durch die automatische Erkennung von Dialogakt-Typen ermöglicht.[18] Lassen sich Dialogakt-Typen automatisch erkennen, so kann auch die Zugehörigkeit von ausgangssprachlichem Input zu einer bestimmten Dialogphase ermittelt werden.

Die Ergebnisse von Teilprojekt 11 ('Semantische Auswertung') lassen erwarten, daß das Konzept des Translationsziels implementierbar und somit für VERBMOBIL auszuwerten ist.

[17] Genaueres zur Begründung dieses Vorschlages findet sich in [VM-Report54].
[18] [s.a. VM-Report28], [VM-Report163].

Für die folgenden Komponenten von VERBMOBIL ergeben sich aus den obigen Ausführungen klare Schlußfolgerungen:

1. Transfer

   Translationsziele ermöglichen eine Spezifikation der Äquivalenzbeziehung. Äquivalenz läßt sich für VERBMOBIL auf die Frage nach der Funktion einer zielsprachlichen Äußerung zurückführen: Was soll mit der Äußerung erreicht werden? Welchen kommunikativen Zweck erfüllt sie? Dieser Zweck fließt in die Definition von Translationszielen ein. Auf dieser Grundlage kann beispielsweise entschieden werden, ob im Falle eines "translation mismatch" (der im Lichte translationswissenschaftlicher Erkenntnisse eher der Normalfall ist) bedeutungsabschwächend oder -verstärkend übersetzt werden soll und ob ggf. weitere Informationen durch eine tiefere Analyse oder durch Rückfrage beim Benutzer beschafft werden müssen.

2. Variable Analysetiefe

   Die Spezifikation von Translationszielen als Adäquatheitskriterium ist auch für Fragen der Architektur im Hinblick auf die Variable Analysetiefe von Bedeutung. Translationsziele können helfen festzulegen, welche Information an welcher Stelle im Translationsprozeß notwendig ist. Der Informationsfluß zwischen einzelnen VERBMOBIL-Komponenten kann damit optimiert werden. Durch das Konzept der Variablen Analysetiefe lassen sich mit Hilfe der Translationsziele Hinweise geben, welche Strategie (Reduktion oder Achievement) jeweils zu wählen ist.

3. Qualitätssicherung

   Wenn definierte Zielvorgaben vorliegen, lassen sich anhand dieser Kriterien Einzelverdolmetschungen im Hinblick auf ihre Adäquatheit überprüfen. Translationsziele geben demnach ein Instrumentarium an die Hand, mittels dessen eine Qualitätskontrolle zu operationalisieren ist (s. Abschnitt 3).

4. Systemtransparenz

   Klare Zielvorgaben machen den NutzerInnen deutlich: Wo liegen die Grenzen der Systemleistung? Welche Erwartungen können in welcher Phase des Dialoges an die Translation gestellt werden?
   Translationsziele geben Anhaltspunkte für eine realistische Einschätzung der Systemleistung von VERBMOBIL und tragen so auch zur Verläßlichkeit des Systems bei. Diese Verläßlichkeit kann die Kooperativität der NutzerInnen erhöhen, sich auf Systemeigenschaften wie z.B. Klärungsdialoge einzulassen.

5 Fazit

Die im vorliegenden Bericht präsentierten Ergebnisse sind in unterschiedlichem Maße in die Arbeiten der anderen Teilprojekte von VERBMOBIL eingeflossen:

- Das Prinzip der Reduktion kann inzwischen als allgemein anerkannt gelten; wieweit auch inhaltliche Information im engeren Sinne reduziert werden soll/darf, ist noch umstritten (das korreliert mit unseren Befragungen bei potentiellen NutzerInnen - wahrscheinlich ist es zweckmäßig, den NutzerInnen nach Möglichkeit eine Auswahl zu erlauben).

- Ambiguitätserhaltende Translationen werden, soweit möglich, im Transfer berücksichtigt.

- Die Ergebnisse zu Defaults und Hypothesenrevision sind bislang nur sehr begrenzt für VERBMOBIL nutzbar - einerseits, weil die menschlichen Strategien in diesem Bereich anscheinend relativ "abenteuerlich" sind - andererseits, weil eine genauere Modellierung zur Zeit auch (noch) nicht vorgesehen ist.

- Das Konzept der Variablen Analysetiefe, verbunden mit dem des Translationsziels, könnte Hinweise darauf geben, wie eine Auswahl zwischen den verschiedenen in VERBMOBIL vorgesehenen Verarbeitungsstrategien automatisch getroffen werden kann. Zu dieser Frage gibt es seit einiger Zeit eine Diskussion mit der Systemgruppe, so daß davon ausgegangen werden kann, daß unsere Überlegungen in die 2. Phase von VERBMOBIL einfließen.

- Die Aufstellung von Bewertungskriterien aus NutzerInnensicht im Zusammenhang mit der Definition von Translationszielen für VERBMOBIL wurde inzwischen mit dem neuen Arbeitspaket 'End-to-end' (Universität Hamburg) diskutiert und wird bei den weiteren Arbeiten zur Evaluation berücksichtigt.

- Unklar ist derzeit, welche unserer Vorschläge zu den Kontrollmöglichkeiten für NutzerInnen verwirklicht werden (sollen); wir halten diesen Aspekt nach wie vor für extrem wichtig.

Es ist also festzuhalten, daß trotz mancher methodischer Schwierigkeiten, die zum großen Teil schon bei der Planung vorhergesehen wurden, einige wesentliche Hypothesen aus der Arbeit des Hildesheimer Teilvorhabens hervorgegangen sind, die sich im VERBMOBIL-Projekt bereits bewähren oder aber demnächst erprobt werden sollen.

Das grundsätzliche Verfahren, nämlich menschliche Strategien und Bewertungen daraufhin zu untersuchen, wie weit sie für ein maschinelles Dolmetschsystem nutzbar sind, hat sich jedenfalls bewährt. Dadurch wurde auch wesentlich dazu beigetragen, eine Brücke zwischen den Forschungsgebieten der Translationswissenschaft und der Maschinellen Übersetzung zu schlagen (dokumentiert in [Hauenschild/Heizmann97]). Diese Bemühungen haben auch international Beachtung gefunden und werden hoffentlich den Weg zu weiteren fruchtbaren Kooperationen ebnen.

6 Im Projekt entstandene Schriften

Die Literaturliste enthält auch einige wenige Schriften von ProjektpartnerInnen, die zwar nicht im Hildesheimer Teilvorhaben entstanden, aber aus der Kooperation hervorgegangen sind.

[**Hauenschild/Heizmann97**]: *Machine Translation and Translation Theory.* Berlin, Mouton deGruyter, 1997.

[**VM-Memo4**] Hauenschild, Christa/Prahl, Birte: *Translationsprobleme - Translationsstrategien.* Verbmobil-Memo 4, Universität Hildesheim, November 1993.

[**VM-Memo13**] Heizmann, Susanne: *Bewertung von Systemen der maschinellen Übersetzung und Anwendbarkeit auf VERBMOBIL.* Verbmobil-Memo 13, Universität Hildesheim, Februar 1994.

[**VM-Memo24**] Bade, Ute/Heizmann, Susanne/Jekat-Rommel, Susanne/Kameyama, Shinichi/Krause, Detlev/Maleck, Ilona/Prahl, Birte/Preuß, Wiebke: *Der Verbmobilsimulator und Wizard of Oz-Experimente in TP 13.* Verbmobil-Memo 24.

[**VM-Memo26**] Heizmann, Susanne/Ahrens, Kerstin : *Bewertung maschineller Dolmetschleistungen - Nutzbare Ansätze aus der Translationskritik.* Verbmobil-Memo 26, Universität Hildesheim, Januar 1995.

[**VM-Memo27**] Heizmann, Susanne/Petzolt, Susanne: *Bewertung von Translationsleistungen - Explorative Vorstudie.* Verbmobil-Memo 27, Universität Hildesheim, November 1994.

[**VM-Memo29**] Mohnhaupt, Corinna: *Maschinelle Uebersetzung und menschliche Translation - Ein Erfahrungsbericht aus Kanada.* Verbmobil-Memo 29, Universität Hildesheim. Juli 1994.

[**VM-Memo34**] Prahl, Birte: *Typen von Reduktion im Translationsprozeß.* Verbmobil-Memo 34, Universität Hildesheim, August 1994.

[**VM-Memo54**] Prahl, Birte: *Menschliche Desambiguierungsstrategien für die Maschinelle Übersetzung? Ein Beispiel für ambiguitätserhaltende Übersetzungen.* Verbmobil-Memo 54, Universität Hildesheim, November 1994.

[VM-Memo58] Petzolt, Susanne/Prahl, Birte: *Menschliche Standardannahmen im VERBMOBIL-Szenario.* Verbmobil-Memo 58, Universität Hildesheim, Dezember 1994.

[VM-Memo73] Susanne Petzolt: *Studie des Dolmetschprozesses im VERBMOBIL-Szenario - Versuchsbeschreibung und Datenmaterial.* Verbmobil-Memo 73, Universität Hildesheim, Mai 1995.

[VM-Memo80] Heizmann, Susanne/Petzolt, Susanne/Prahl, Birte: *Klassifikation von Translationsproblemen und Auswertung von Bewertungskriterien im Hinblick auf den Demonstrator.* Verbmobil-Memo 80, Universität Hildesheim, Mai 1996.

[VM-Memo88] Krause, Detlev: *Akzeptanz gegenüber VERBMOBIL (Prioritätenliste II).* Verbmobil-Memo 88, Universität Hamburg, 1995.

[VM-Memo97] Petzolt, Susanne: *Hypothesenrevision im Translationsprozeß - Versuchsbeschreibung und Datenmaterial.* Verbmobil-Memo 97, Universität Hildesheim, November 1995.

[VM-Memo98] Heizmann, Susanne/Petzolt, Susanne: *Bewertung maschineller Dolmetschleistungen - anhand der Videoaufzeichnung einer (simulierten) Nutzungssituation des Systems VERBMOBIL.* Verbmobil-Memo 98, Universität Hildesheim, November 1995.

[VM-Memo121] Achterfeld, Silke/Hauenschild, Christa/Icking, Barbara/Wöllbrink, Birgit: *Beschreibung von menschlichen Desambiguierungsstrategien im Hinblick auf verschränkte Ambiguitäten.* Verbmobil-Memo 121, Universität Hildesheim, Juni 1997.

[VM-Report28] Schmitz, Birte/Jekat-Rommel, Susanne: *Eine zyklische Approximation an Sprechhandlungstypen - zur Annotierung von Äußerungen in Dialogen.* Verbmobil-Report 28, TU Berlin. September 1994.

[VM-Report53] Schömann, Munira: *Menschliche Informationsverarbeitungsprozesse bei der Disambiguierung.* Verbmobil-Report 53, Universität Hildesheim, Januar 1995.

[VM-Report54] Prahl, Birte/Petzolt, Susanne/Heizmann, Susanne/ Hauenschild, Christa: *Variable Analysetiefe und Bewertungskriterien in VERBMOBIL: Translationswissenschaftliche Grundlagen.* Verbmobil-Report 54, Universität Hildesheim, Februar 1995.

[VM-Report94] Krause, Detlev/Bade, Ute/Preuß, Wiebke: *Die Experimente mit dem VERBMOBIL-Simulator. Design - Ablauf - Daten.* Verbmobil-Report 94, Universität Hamburg, September 1995.

[VM-Report163] Schmitz, Birte: *The Translation Objective in Automatic Dialogue Interpreting.* Verbmobil-Report 163, Technische Universität Berlin, August 1996.

[VM-Report197] Prahl, Birte/Petzolt, Susanne: *Translation problems and translation strategies involved in human and machine translation: Empirical studies.* Verbmobil-Report 197, Universität Hildesheim, Juni 1997.