Feature

„Maschinelle Übersetzung Japanisch-Deutsch: Stand der Technik und Ausblick“

Werner Winiwarter

Die Kenntnis der japanischen Sprache ist der Schlüssel für ein besseres Verständnis der japanischen Kultur und Gesellschaft, die uns sonst oft so fremdartig und widersprüchlich scheint. Ganz egal, ob man sich im Land befindet oder aus der Ferne mehr über dieses faszinierende Land in Erfahrung bringen möchte, eröffnet erst die japanische Sprache viele Einsichten, die einem ansonsten verborgen bleiben.

In den letzten Jahren hat die Anzahl der japanischen Dokumente, die über das Internet weltweit verfügbar sind, rasant zugenommen. Nach Spanisch und Deutsch ist Japanisch die Sprache mit der höchsten Zuwachsrate im Internet und gemessen an der Zahl an Internetbenutzern liegt Japan gar an zweiter Stelle hinter den USA. Das Internet bietet also einen riesigen Schatz an Wissen über japanische Kultur, Gesellschaft, Politik, Wirtschaft und viele andere Bereiche.

Darüber hinaus ist das Internet auch ein wichtiges Instrument für alle, die über das Studium japanischer Texte ihre Japanischkenntnisse verbessern wollen. Im allgemeinen bietet das Lesen von fremdsprachlichen Dokumenten eine ausgezeichnete Möglichkeit, neue Vokabel, Formulierungen und grammatikalische Konstruktionen in ihrem natürlichen Kontext aus dem Zusammenhang heraus auf relativ einfache Weise zu erfassen und zu erlernen. Während diese Form des Spracherwerbs für viele Sprachen sehr gut funktioniert, stellt sich einem im Fall von Japanisch das komplizierte Schriftsystem in den Weg. Die japanische Schrift besteht bekanntlich aus den zwei Silbenschriften Hiragana und Katakana sowie den chinesischen Schriftzeichen Kanji. Während die Silbenschriften mit jeweils 46 Zeichen noch relativ einfach zu erlernen sind, gibt es mehrere tausend, meist sehr komplex aufgebaute Kanji. Eine weitere große Schwierigkeit beim Lesen japanischer Texte ist die Tatsache, daß die einzelnen Worte nicht durch Zwischenräume getrennt sind, sodaß man zunächst die einzelnen Wortgrenzen erraten muß. Wenn man an einer unverständlichen Textstelle hängen bleibt, muß man zunächst herausfinden, wo ein unbekanntes Wort beginnt, bevor man dieses in einem Wörterbuch nachschlagen kann.

Allerdings kann man dafür nur dann ein herkömmliches zweisprachiges Wörterbuch verwenden, wenn man die Aussprache des Wortes kennt. Da leider Kanji abhängig vom Kontext über mehrere Aussprachen oder Lesungen verfügen können, muß man oftmals ein Kanji-Lexikon zu Rate ziehen. In einem Kanji-Lexikon sind die Kanji nach 214 Grundelementen oder Radikalen eingeteilt, wobei die Suche nach einzelnen Kanji noch dadurch erschwert wird, daß die Radikale unterschiedliche Erscheinungsformen je nach Position im Kanji aufweisen können. Alles in allem bedeutet dies, daß das Lesen, Verstehen und Übersetzen japanischer Texte mit traditionellen Hilfsmitteln eine oftmals frustrierende und langwierige Aufgabe darstellt, die schon manchen Japanologiestudenten zur Verzweiflung getrieben hat.

Glücklicherweise bieten Computer nicht nur die Möglichkeit, japanische Dokumente über das Internet zu recherchieren, sondern es existiert auch bereits eine Vielzahl an nützlichen Werkzeugen, um dem Benutzer beim Lesen von japanischen Texten unter die Arme zu greifen. In diesem Zusammenhang sind insbesondere die Initiativen EDICT von Jim Breen und WaDokuJT von Ulrich Apel hervorzuheben. EDICT ist ein frei verfügbares Japanisch-Englisch-Lexikon mit über 100.000 Einträgen, während WaDokuJT als frei erhältliches Japanisch-Deutsch-Lexikon mit zur Zeit 206.820 Einträgen speziell für deutschsprachige Benutzer von großem Nutzen sein dürfte (alle Links sind am Ende des Artikels aufgelistet). Ebenfalls von Jim Breen entwickelt und kostenlos zur Verfügung gestellt wurde das Kanji-Lexikon KANJIDIC, das umfangreiche Namenslexikon ENAMDICT sowie eine Reihe weiterer interessanter lexikalischer Ressourcen. Alle diese Lexika werden laufend erweitert und verbessert, wobei hierfür wiederum freiwillige Beiträge aus der ganzen Welt gesammelt werden. Es ist dies ein sehr schönes Beispiel dafür, daß durch das Zusammenarbeiten an einer gemeinsamen Aufgabe und die Bereitstellung der Ergebnisse in Summe für alle langfristig mehr erreicht werden kann als durch kurzsichtiges Revierdenken. Diese Lexika sind auch die Grundlage für viele weiterführende Werkzeuge zur Unterstützung beim Lesen japanischer Texte. Dies reicht von einfachen Suchschnittstellen über japanische Texteditoren bis hin zu Websites, die japanische Webseiten analysieren und mit Zusatzinformation hinterlegen. Ein Beispiel für letztere sind POPjisho und Rikai, bei denen man nur mit der Maus auf eine bestimmte Textstelle zeigen muß, um Informationen über Kanji sowie englische oder deutsche Wortbedeutungen in einem Pop-up-Fenster angezeigt zu bekommen. Obwohl diese Hilfsmittel sehr hilfreich sind, gibt es doch oftmals noch Schwachstellen in Bezug auf die korrekte Zerlegung in einzelne Worte sowie das Auffinden von Informationen für konjugierte Wortformen.

Auch in meiner persönlichen Forschungsarbeit habe ich mich seit vielen Jahren mit der Entwicklung von Arbeitsumgebungen für die Unterstützung von

Sprachstudenten und anderen Interessierten beim Lesen, Bearbeiten und Übersetzen von japanischen Texten auseinandergesetzt. Während frühe Arbeiten noch auf eine Umgebung zur flexiblen Anreicherung von Webseiten mit zusätzlichen Informationen ausgelegt waren (Winiwarter, 1999), hat sich in den letzten Jahren meine Aufmerksamkeit auf eine Arbeitsumgebung konzentriert, die in übliche Microsoft Office Anwendungen eingebettet werden kann, insbesondere in den Texteditor Microsoft Word. Dies ermöglicht nicht nur das passive Konsumieren von japanischen Dokumenten, sondern auch das aktive Arbeiten mit einem Text in einer vertrauten Arbeitsumgebung. Unter Rückgriff auf WaDokuJT wurde von mir die korrekte Segmentierung in einzelne Worte, die Lexikonsuche nach konjugierten Wortformen sowie die flexible Hinzufügung eigener Lexikoneinträge implementiert.

Obwohl erste Erfahrungen mit dem Einsatz dieser Lernumgebung zeigten, daß sie bei Sprachstudenten sehr positiven Anklang fand, mußte ich allerdings auch feststellen, daß lexikalische Informationen auf Wortebene in vielen Fällen einfach nicht ausreichend sind, um die korrekte Bedeutung eines japanischen Texts zu erfassen. Der Hauptgrund hierfür liegt sicherlich in der Komplexität der Übersetzungsaufgabe für das Sprachpaar Japanisch-Deutsch, welche durch die stark abweichenden Grammatiken bedingt ist. Deutsch verfügt über ein ausgeprägtes Flexionssystem mit zahlreichen Deklinationen und Konjugationen, um syntaktische Eigenschaften, wie Zahl, Zeitstufe, Aussageweise oder Handlungsrichtung auszudrücken. Im Gegensatz dazu ist Japanisch hinsichtlich vieler dieser Merkmale mehrdeutig, z.B. gibt es im Japanischen keine Artikel, um Geschlecht anzugeben, keine Deklination für die Bestimmung der Zahl oder des Falls sowie nur zwei Zeitstufen. Diese Situation wird noch dadurch drastisch verschlimmert, daß Japaner oftmals Satzteile einfach weglassen, wenn diese aus anderen Informationen abgeleitet werden können, was speziell auch für das Subjekt eines Satzes gilt.

All dies spricht dafür, zusätzliche Hilfen anzubieten, die über die Übersetzung einzelner Worte hinausgehen. Aus diesem Grund begannen sich meine Forschungsarbeiten in letzter Zeit immer stärker in Richtung maschinelle Übersetzung zu verlagern. Nach ernüchternden Erfahrungen mit existierenden Systemen und ersten erfolglosen Implementierungsversuchen basierend auf bestehenden Ansätzen wurde mir bald klar, daß ich methodisch völliges Neuland betreten müßte, um für diese intellektuelle Herausforderung eine praktikable Lösung erarbeiten zu können. Die wichtigste Anforderung an mein Übersetzungssystem war, daß das System in der Lage ist, sein gesamtes Übersetzungswissen aus Beispielübersetzungen zu lernen, d.h. daß keine manuell erzeugten Übersetzungsregeln in das System eingegeben werden müssen. Weiterhin sollte das System in der Lage sein, adaptiv sein Wissen an die individuellen Vorlieben eines Benutzers anzupassen. Einfach ausgedrückt,

wenn mir eine Übersetzung des Systems nicht gefällt, korrigiere ich sie einfach und das System berücksichtigt diese Änderung bei zukünftigen Übersetzungen. Aufgrund dieser Flexibilität eignet sich mein Übersetzungssystem optimal für den Einsatz im Sprachunterricht, da das System mit dem Sprachstudenten mitlernt und einen aktiven bidirektionalen Wissensaustausch fördert. Langweilige Lese- und Übersetzungsaufgaben werden somit zu interessanten interaktiven Experimenten. Um Sprachstudenten auch ein entsprechend komfortables Arbeiten zu ermöglichen, habe ich das Übersetzungssystem mit der bestehenden Arbeitsumgebung für Microsoft Word integriert, sodaß auf alle Übersetzungsfunktionen direkt aus dem Texteditor heraus zugegriffen werden kann. Der daraus resultierenden Lernumgebung für computergestützten Spracherwerb habe ich den Namen PETRA (Personal Embedded Translation and Reading Assistant) gegeben.

Der restliche Artikel ist so aufgebaut, daß ich zunächst einen Überblick über den Stand der Technik auf dem Bereich der maschinellen Übersetzung für Japanisch-Deutsch gebe, bevor ich meinen eigenen Ansatz vorstelle. Schließlich werde ich kurz zeigen, wie das Übersetzungssystem in die Lernumgebung PETRA integriert wurde, und über erste Erfahrungen beim praktischen Einsatz von PETRA berichten.

Stand der Technik für maschinelle Übersetzung Japanisch-Deutsch

Die Forschung und Entwicklung auf dem Gebiet der maschinellen Übersetzung hat allgemein bereits eine sehr lange Geschichte. Die ersten großangelegten Versuche wurden schon in den 50er Jahren zur Zeit des Kalten Kriegs für militärische und Spionagezwecke durchgeführt, d.h. für die Übersetzung sensitiver russischer Dokumente ins Englische. Was wahrscheinlich symptomatisch für die gesamte Thematik der Sprachtechnologie ist, fand auch damals bereits statt, nämlich daß die Komplexität der Aufgabe vollkommen unterschätzt wurde. Man dachte sich, Übersetzung sei ein reines Codierungsproblem und könnte mit (für die damalige Zeit) schnellen Rechnern rasch gelöst werden. Die Ergebnisse waren allerdings vernichtend. Diese Abfolge – überzogene Erwartungen, fehlendes Problemverständnis, voreilige Versprechungen, enttäuschende Resultate, frustrierte Anwender, desillusionierte Auftraggeber – zieht sich durch die gesamte Geschichte der maschinellen Übersetzung und vieler anderer Bereiche der Sprachtechnologie bis hin zur Gegenwart.

Der heutige Stand der Technik für maschinelle Übersetzung sieht so aus, daß es recht brauchbare Lösungen gibt, allerdings nur für sehr eingeschränkte Anwendungsbereiche mit kleinem Vokabular. Ein bekanntes, oft zitiertes Beispiel ist das System METEO für die automatische Übersetzung von

Wetterberichten. Wenn man an eine allgemeinere Verwendung denkt, dann bieten nur sehr ähnliche Sprachpaare die Aussicht auf ein Übersetzungsergebnis, das akzeptabel oder zumindest verständlich ist. Es ist die allgemein vorherrschende Lehrmeinung, daß vollautomatische hochqualitative Übersetzung ohne thematische Einschränkungen und ohne menschliche Einflußnahme weit außerhalb der Möglichkeiten heutiger maschineller Übersetzungstechnologien liegt. Es bestehen sogar berechtigte Zweifel, ob dieses Ziel jemals erreicht werden wird (Hutchins, 2003).

Diese ernüchternde Bewertung trifft sicher auf transferbasierte Übersetzungssysteme zu, welche den Großteil der kommerziell verfügbaren Systeme ausmachen. Beim transferbasierten Ansatz wird versucht, Übersetzungsregeln für ein bestimmtes Sprachpaar aufzustellen. Diese sogenannten Transferregeln werden zumeist manuell von Experten in jahrelanger Kleinarbeit zu einer umfangreichen Regelbasis zusammengefügt. Das Hauptproblem bei diesem Ansatz ist, daß mit zunehmender Größe der Regelbasis es immer schwieriger wird, diese in einem konsistenten Zustand zu erhalten. Jede neue Regel, die hinzugefügt wird, kann unerwünschte Auswirkungen auf eine Vielzahl anderer Regeln haben, sodaß Sätze, die vorher problemlos korrekt übersetzt wurden, plötzlich zu falschen Ergebnissen führen. Somit muß für jede neue Regel ein komplizierter Konfliktauflösungsprozeß gestartet werden, der auch wiederum Umformulierungen zahlreicher anderer existierender Regeln bewirken kann. Kurz zusammengefaßt, große Regelbasen erhalten mehr und mehr einen statischen Charakter, da Erweiterungen immer schwieriger werden.

Diese Problematik trifft in noch viel größerem Maße auf interlinguabasierte Systeme zu. Diese setzen sich das ambitionierte Ziel, eine sprachunabhängige Zwischenrepräsentation (die Interlingua) zu finden, die zwischen einer beliebigen Anzahl von Sprachen vermittelt. Eine solche Interlingua hätte natürlich den großen Vorteil, daß ich eine viel kleinere Anzahl an Übersetzungskomponenten bauen müßte, da ich z.B. nur einmal von Japanisch in die Interlingua und dann von der Interlingua in jede beliebige Sprache übersetzen könnte. Das klingt alles zu schön, um wahr zu sein, und tatsächlich gibt es natürlich in der praktischen Umsetzung unüberwindbare Schwierigkeiten, da jede sprachliche Nuancierung für jede einzelne Sprache in dieser Interlingua abbildbar sein müßte (man denke nur an das in der Linguistik beliebte Beispiel der zahlreichen Worte für Schnee im Grönländischen). Dies führt sehr rasch zu einer kombinatorischen Explosion, sodaß der praktische Einsatz einer solchen Interlingua für die Übersetzung allgemeiner Texte für eine größere Anzahl von Sprachen leider nur ein unerfüllbarer Traum ist.

Wenn man sich die Weiterentwicklung der Übersetzungsqualität in den letzten 10 Jahren ansieht, kommt man zu der ernüchternden Einsicht, daß es bei den

Übersetzungsergebnissen kaum zu nennenswerten Verbesserungen gekommen ist (Somers, 2003). Ein Hauptgrund dafür ist sicher der bereits angesprochene statische Charakter der manuell erstellten Wissensbasen. Wenn ein Benutzer ein solches System erwirbt, dann muß er die gelieferten Resultate so akzeptieren, wie sie sind, oder sich nach einem anderen System umsehen. All der stundenlange intellektuelle Aufwand, den weltweit Benutzer in die Überarbeitung solcher mangelhafter Übersetzungen investieren müssen, geht üblicherweise wieder verloren und fließt nicht in zukünftige Übersetzungen ein.

Als ein möglicher Ausweg aus diesem Dilemma wurden korpusbasierte Ansätze vorgeschlagen. Diese arbeiten auf der Grundlage großer bilingualer Textsammlungen oder Korpora. Ein allgemeines Problem hierbei ist, daß meistens Rohtexte alleine nicht für den Einsatz als Datenbasis in Sprachtechnologieanwendungen ausreichen, sondern daß die Texte manuell oder zumindest semiautomatisch mit zusätzlichen Daten (Wortkategorien, Satzstrukturen, etc.) angereichert werden müssen, damit sie sinnvoll angewendet werden können. Für maschinelle Übersetzung kommt noch die große Schwierigkeit hinzu, daß Quell- und Zieltexte exakt aufeinander ausgerichtet werden müssen, um daraus Beispielübersetzungen für Sätze oder Formulierungen ableiten zu können. Dies setzt natürlich eine hochqualitative und originalgetreue Übersetzung voraus. Sobald größere Textpassagen näherungsweise übersetzt, Sätze zusammengefaßt, Teile weggelassen oder neu angeordnet werden, scheitern meistens alle Versuche eines automatisierten Abgleichs. Je umfangreicher der bilinguale Korpus, desto schwieriger wird es auch, die Qualität des Datenmaterials zu gewährleisten. In Bezug auf die Verfügbarkeit solcher Korpora für Japanisch ist auch anzumerken, daß umfangreiche Anstrengungen für das Erstellen von Korpora für Japanisch-Englisch unternommen wurden, während es leider in Hinsicht auf Japanisch-Deutsch kaum nennenswertes Material gibt.

Hat man allerdings doch einen brauchbaren Korpus für ein Sprachpaar zur Verfügung, dann kann man unter Anwendung statistischer Übersetzungstechniken (Brown, 1990) versuchen, auf der Grundlage der Häufigkeiten des Auftretens im Korpus für die einzelnen Worte in einem Satz die wahrscheinlichsten Übersetzungen zu erraten. Prinzipiell wird also jedes Wort einzeln übersetzt und dann versucht, die Worte wieder in eine richtige Reihenfolge zu bringen. Wie man sich vorstellen kann, funktioniert ein solcher statistischer Ansatz in Reinform wiederum nur für sehr ähnliche Sprachen halbwegs brauchbar.

Für Japanisch hat in den letzten Jahren in der Forschung auf dem Bereich der Verwendung von Korpora für maschinelle Übersetzung eher der beispielbasierte Ansatz (Sato, 1991) Zulauf bekommen. Hierbei ist die Grundidee, aus einem bilingualen Korpus Übersetzungsbeispiele für Formulierungen zu sammeln und

dann für die einzelnen Teile eines neuen Satzes in dieser Beispielsammlung nach den ähnlichsten Beispielen zu suchen. Die in der Beispielsammlung gefundenen Übersetzungen für die Satzteile werden dann wiederum zum vollständigen Satz in der Zielsprache zusammengebaut. Mittlerweile gibt es eine Vielzahl von Arbeiten zu dieser Forschungsrichtung, die sich meist darin unterscheiden, wie sie die Übersetzungsbeispiele in der Beispielsammlung repräsentieren. Während manche Ansätze strukturelle Darstellungen für alle konkreten Beispiele speichern, wird in anderen Systemen versucht, aus ähnlichen Beispielen verallgemeinerte Beispiele abzuleiten, in denen dann konkrete Worte durch Variable ersetzt werden. All diesen Aktivitäten haftet allerdings der große Nachteil an, daß fast alle Repräsentationen für Übersetzungsbeispiele in der Beispielsammlung für ein beispielbasiertes System vernünftiger Größe wiederum entweder manuell erstellt oder zumindest auf Korrektheit überprüft werden müssen. Dies relativiert also somit die Sinnhaftigkeit des beispielbasierten Ansatzes für praktische Anwendungen.

Um diesen Überblick über die verschiedenen Ansätze zu einem Abschluß zu bringen, möchte ich das Resümee ziehen, daß man sich bei der Entwicklung eines neuen Übersetzungssystems unter Rückgriff auf existierende Techniken zwischen zwei Möglichkeiten entscheiden kann. Man kann mehrere Personenjahre Aufwand in die manuelle Erstellung von Transferregeln oder einer Interlingua investieren und hat am Ende eine große statische Wissensbasis, die kaum noch wartbar ist, oder man setzt sein Vertrauen auf statistische Verfahren basierend auf umfangreichen bilingualen Korpora, muß dann aber mangelhafte Übersetzungen in Kauf nehmen, die durch die ungenaue approximative Sprachmodellierung bedingt sind. Beispielbasierte Verfahren bieten gewissermaßen einen Mittelweg an. Man kann sich hierbei aussuchen, in welchem Ausmaß Personalressourcen in die Erweiterung und Korrektur der Beispielsammlung fließen sollen, um die Qualität des Systems zu heben.

Wem diese Darstellung zu negativ gefärbt erscheint, dem möchte ich in Abb. 1 die Ergebnisse eines unterhaltsamen Experiments nahe legen. Ich habe eine Reihe von maschinellen Übersetzungssystemen, die frei im Web verfügbar sind, für einen sehr einfach gehaltenen japanischen Satz ohne Spezialvokabel abgefragt und die Übersetzungsergebnisse mitnotiert. Leider liefert nur das erste angeführte System eine Übersetzung ins Deutsche, alle anderen sind ausschließlich für die Übersetzung ins Englische ausgelegt. Selbst bei so einfachen Sätzen werden gerade einmal die Worte annähernd korrekt oder zumindest ähnlich übersetzt, beim Zusammensetzen der Worte in einen sinnvollen Satzzusammenhang herrscht allerdings schon hier Ratlosigkeit. Besonders amüsant ist z.B. die Interpretation, daß das Buch zum Mittelalter wurde, und warum das Buch gewaschen worden sein soll, ist auch nicht wirklich nachvollziehbar. Es ist jedenfalls offensichtlich, daß solche

Übersetzungsergebnisse nicht als zufriedenstellend bezeichnet werden können. Sie sind weit entfernt von hochqualitativen Übersetzungen, ja manchmal ist es sogar schwierig oder unmöglich, überhaupt die Bedeutung des ursprünglichen Satzes aus den verstümmelten Übersetzungen zu rekonstruieren.

Japanischer Satz:
いまのような形の本は、中世になって、はじめてあらわれた。
Transkription:
ima no you na katachi no hon wa, chuusei ninatte, hajimete arawareta.
Menschliche Übersetzung:
Das Buch in seiner heutigen Form ist im Mittelalter zum ersten Mal aufgetreten.
The book in its present form appeared for the first time in the Middle Ages.
Maschinelle Übersetzung durch WorldLingo (www.worldlingo.com/products_services/worldlingo_translator.html):
Das Buch der Form die gegenwärtige Weise erschien und war mittleres Alter, zum ersten Mal.
The book of the shape the current way appeared, being Middle Ages, for the first time.
Maschinelle Übersetzung durch Excite (www.excite.co.jp/world/url/):
The book of a form like now appeared only after it became medieval times.
Maschinelle Übersetzung durch TransLand (www.brother.co.jp/jp/honyaku/demo/index.html):
The books of a form like the present became medieval times, and it was begun and washed.
Maschinelle Übersetzung durch linguatec (http://www.linguatec.net/online/ptwebtext/index_en.shtml):
The book which is like a now of shape became the Middle Ages, began and appeared..

Abb. 1: Beispiele für maschinelle Übersetzungssysteme

Leider habe ich mit kommerziell in Japan vertriebenen Übersetzungssystemen ähnlich schlechte Erfahrungen machen müssen. Hinzu kommt bei solchen Produkten noch die Problematik, daß diese natürlich für japanische Benutzer und fast ausschließlich für die Übersetzung ins Englische ausgelegt sind. Sie laufen daher oft nur auf japanischen Betriebssystemen, verfügen über eine japanische Benutzeroberfläche und werden von Dokumentation und Hilfetexten auf Japanisch begleitet. Die meisten dieser Systeme (z.B. ATLAS) bieten eingeschränkte Möglichkeiten, die Form der Ausgaben zu beeinflussen, z.B. eher Verwendung aktiver oder passiver Satzkonstruktionen. Wenn man allerdings in der Lage ist, sich durch diese komplexen Auswahlmenüs durchzukämpfen, braucht man wahrscheinlich auch kein Übersetzungssystem mehr. Ein weiterer Schwachpunkt bei vielen Systemen ist, daß diese als isolierte Insellösungen mit eigenen Editoren konzipiert wurden, während es nur selten möglich ist, direkt auf die Übersetzungsfunktionalität von der gewohnten Arbeitsumgebung aus (z.B. Microsoft Word) zuzugreifen. Eine schwerwiegende

Problematik, nämlich die der unterschiedlichen Codierungssysteme für japanische Texte (JIS, Shift-JIS, EUC) und der damit einhergehenden Konvertierungsprobleme, ist zumindest bei neueren Systemen durch die durchgängige Verwendung von Unicode als weltweitem Standard für Textcodierung entschärft worden. Der Vollständigkeit halber wären noch Arbeitsumgebungen für professionelle Übersetzer (wie z.B. TRADOS) zu erwähnen, die Terminologieunterstützung sowie einen Übersetzungsspeicher anbieten, um auf in der Vergangenheit durchgeführte Übersetzungen zugreifen zu können. Allerdings bieten solche Umgebungen keinerlei automatische Übersetzungsfunktionalität und sind somit eher nur für einen stark eingeschränkten Personenkreis von Nutzen.
Kurz zusammengefaßt gibt es also maschinelle Übersetzungssysteme, die man käuflich erwerben kann, allerdings nur für japanische Benutzer für die Zielsprache Englisch und das mit enttäuschenden Ergebnissen. Für deutschsprachige Benutzer, die japanische Texte lesen, verstehen, bearbeiten und übersetzen möchten, sind diese leider nur von minimalem Nutzen. Insbesondere deutschsprachige Japanologiestudenten benötigen vielmehr eine integrierte Lernumgebung, die sich ihrem Lernfortschritt laufend anpaßt, d.h. mit dem Studenten mitlernt und situationsbezogen die benötigte Unterstützung anbietet.

Inkrementelles Erlernen von Transferregeln aus Übersetzungsbeispielen

In dem Ansatz zur maschinellen Übersetzung, den ich im Laufe der letzten Jahre entwickelt habe, verwende ich Übersetzungsbeispiele des Benutzers, um daraus automatisch neue Transferregeln zu erlernen. Von der Bedienung her entspricht das dem Speichern von Übersetzungsbeispielen in einem Übersetzungsspeicher. Der große Unterschied zu solchen Systemen ist allerdings, daß ich nicht einfach den japanischen und deutschen Satz für eine Übersetzung speichere, sondern für beide Sätze die Satzbäume aufbaue und durch einen strukturellen Abgleich der beiden Satzbäume neue Transferregeln ableite. Ich erhalte also abstrahiertes und verallgemeinertes Übersetzungswissen, das ich dann sofort für neue Übersetzungen anwenden kann.
Abbildung 2 zeigt schematisch die Vorgehensweise beim Erlernen neuer Regeln. Jedes Satzpaar wird zunächst von den Lemmatisierungsmodulen analysiert. Diese erzeugen eine korrekte Segmentierung der Sätze in einzelne Worte. Die Worte werden dabei lemmatisiert, d.h. flektierte Wortformen auf ihre Grundform gebracht und mit Informationen über die Wortart und andere syntaktische Eigenschaften (z.B. Zahl oder Zeitstufe) ergänzt. In Abb. 3 ist zum besseren Verständnis die deutsche Wortliste für den Beispielsatz aus Abb. 1 ersichtlich.

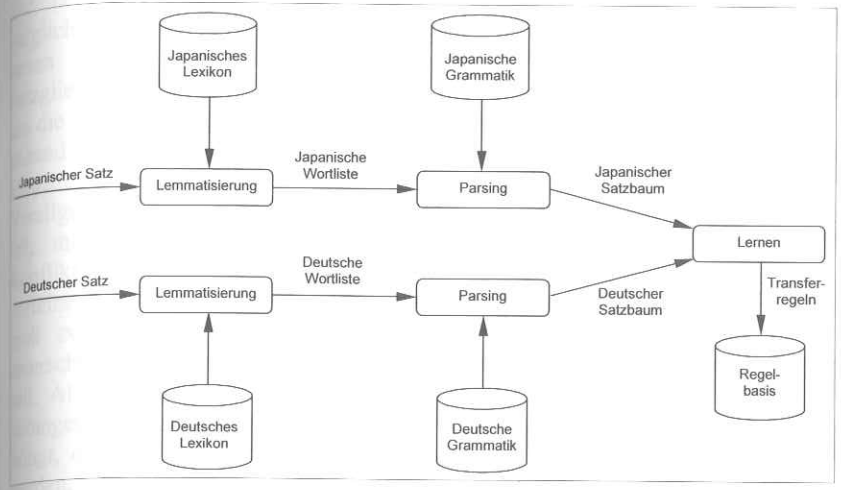

Abb. 2: Lernmodus

Für Japanisch ist das Erstellen dieser Wortliste ungleich schwerer als für Deutsch, da bedingt durch die fehlenden Zwischenräume zwischen den einzelnen Worten zunächst die Aufgabe der korrekten Segmentierung gelöst werden muß. Der japanische Satz wird hierfür als eine lange Zeichenkette behandelt, von der jeweils der Anfang mit möglichen Worten im Lexikon verglichen wird. Diese Suche wird durch die zahlreichen Konjugationsformen für Verben und Adjektive im Japanischen erschwert, da alle Kombinationen von Wortstämmen und Endungen für diese beiden Wortklassen ausprobiert werden müssen. Hat man das richtige Wort für den Satzanfang gefunden, entfernt man es von der Zeichenkette und beginnt die Suche erneut für das nächste Wort. Man kann sich vorstellen, daß dies bei einem großen Lexikon eine große Anzahl an Vergleichen mit sich bringt, sodaß die Segmentierung hohe Anforderungen an die Leistungsfähigkeit der verwendeten Softwarearchitektur stellt.

Deutscher Satz:
Das Buch in seiner heutigen Form ist im Mittelalter zum ersten Mal aufgetreten.

das/art	Artikel
Buch/nsg	Nomen Singular
in/prp	Präposition
sein/ppr	Possessivpronomen
heutig/apo	Adjektiv Positiv
Form/nsg	Nomen Singular
sein/hps	Hilfsverb Präsens
in/kon	Kontraktion
Mittelalter/nsg	Nomen Singular
zu/kon	Kontraktion
erst/ord	Ordinalzahl
Mal/nsg	Nomen Singular
auftreten/vpp	Verb Partizip Perfekt
./pun	Punkt

Abb. 3: Beispiel einer deutschen Wortliste

Sind die beiden Wortlisten ermittelt, ist der nächste Schritt, die Satzstrukturen zu analysieren. Diese Aufgabe übernehmen die *Parsingmodule*, welche die Wortlisten unter Anwendung von Grammatikregeln in Satzbäume umwandeln. Ein interessantes Detail ist hierbei, daß im Japanischen ausschließlich Postpositionen verwendet werden, die ja im Deutschen nur selten vorkommen (z.B. „den Weg entlang"). Da zusätzlich das Prädikat immer am Ende des Satzes steht, ist es einfacher, einen japanischen Satz von rechts nach links zu parsen. Abbildung 4 zeigt den Satzbaum für unseren japanischen Beispielsatz. Wie ersichtlich ist, wird die Information über konjugierte Wortformen getrennt angegeben, z.B. drückt die *ta*-Form die Vergangenheit aus.

いまのような形の本は、中世になって、はじめてあらわれた。
ima no you na katachi no hon wa, chuusei ninatte, hajimete arawareta.
Das Buch in seiner heutigen Form ist im Mittelalter zum ersten Mal aufgetreten.

bzw	ver	あらわれる						Beziehungswort – Verb – arawareru – auftreten
bzf	vta							Beziehungswortform – ta-form
pav	adv	はじめて						prädikatives Adverb – Adverb – hajimete – zum ersten Mal
abe	bzw	nom	中世					adverbiale Bestimmung – Beziehungswort – Nomen – chuusei – Mittelalter
	php	par	になって					Phrasenpartikel – Partikel – ninatte – in
sub	bzw	nom	本					Subjekt – Beziehungswort – Nomen – hon – Buch
	anp	bzw	nom	形				attributive Nominalphrase – Beziehungswort – Nomen – katachi – Form
		anp	bzw	nom	いま			attributive Nominalphrase – Beziehungswort – Nomen – ima – heutig

Abb. 4: Beispiel eines japanischen Satzbaums

Die beiden Satzbäume werden schließlich vom *Lernmodul* miteinander verglichen, um neue Transferregeln abzuleiten. Wir starten unsere Suche nach neuen Regeln auf der Satzebene, bevor wir nach übereinstimmenden Satzgliedern (Subjekt, Objekt, adverbiale Bestimmungen, etc.) Ausschau halten, um die Suche rekursiv für diese Satzglieder fortzusetzen. Jede neue Regel wird anhand der existierenden Regeln in der Regelbasis überprüft, ob nicht durch das Hinzufügen der Regel ein Konflikt entsteht. Da ich bestmögliche Verallgemeinerung anstrebe, um die Regelbasis kompakt zu halten, versuche ich, möglichst allgemeingültige Regeln abzuleiten. Die Auflösung von Konflikten beruht daher darauf, widersprüchliche allgemeine Regeln solange zu verfeinern, bis alle Regeln wieder zu korrekten Übersetzungen führen. Hierbei muß zwischen Regeln für allgemeine Fälle und Regeln für Ausnahmen unterschieden werden können, damit man weiß, welche Regel verfeinert werden soll. Als Grundlage für diese Entscheidung wird die Auftretenshäufigkeit in vorangegangenen Übersetzungen herangezogen, was den wesentlichen Vorteil bringt, daß es keine Rolle spielt, in welcher Reihenfolge die Regeln gelernt werden. Das Lernmodul hat also kein Problem damit, wenn zuerst ein Satz mit einer Ausnahme übersetzt wird und erst später der allgemeingültige Fall auftritt. In Abb. 5 wird gezeigt, wie Transferregeln aus einem Übersetzungsbeispiel gelernt werden. Ich gehe hierbei nicht auf die interne Darstellung der Regeln ein, für technisch versierte Leser verweise ich auf Winiwarter, 2004. Die gestrichelten Linien kennzeichnen Regeln, die für die Übersetzung des konkreten Beispiels eigentlich nicht notwendig sind, aber trotzdem gelernt werden, um so viele Informationen wie möglich aus dem Beispiel zu extrahieren. Die Bestimmtheit und der Numerus für das Subjekt stimmen mit den Defaultwerten „bestimmt" und „Singular" überein, sodaß keine neue Regel für diese beiden syntaktischen Eigenschaften gelernt werden muß.

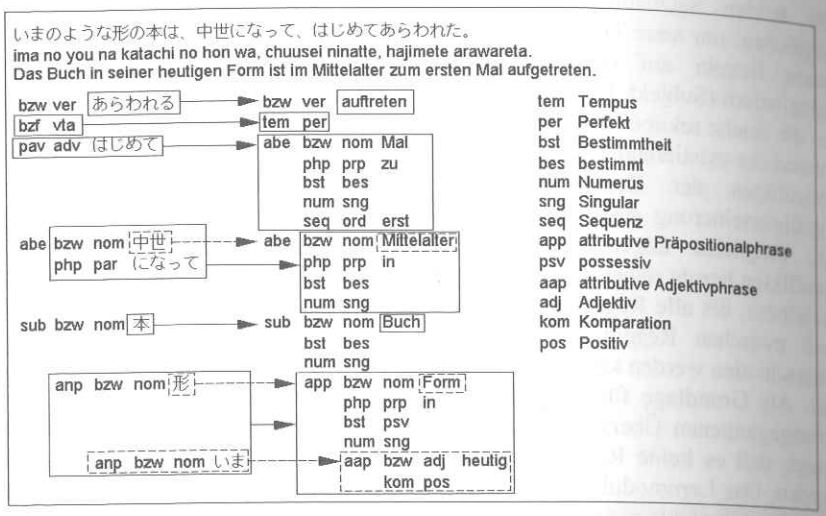

Abb. 5: Beispiel für gelernte Transferregeln

Für die Anwendung der gelernten Transferregeln auf die Übersetzung eines neuen Satzes wird folgendermaßen vorgegangen (siehe Abb. 6). Zuerst wird der Satz lemmatisiert und geparst, um den japanischen Satzbaum zu erzeugen. Danach wird das *Transfermodul* aufgerufen, welches den Satzbaum von oben nach unten durchwandert und in der Regelbasis nach Transferregeln sucht, die angewendet werden können. Die Regeln sind dabei so flexibel definiert, daß eine Regel auch nur bestimmte Teile eines Satzglieds verändern kann, während alle anderen Teile unverändert bleiben, um durch zusätzliche Regeln in darauffolgenden Arbeitsschritten transformiert zu werden. Somit muß das Transfermodul in der Lage sein, effektiv auf einem Satzbaum zu arbeiten, der aus einer Mischung von japanischen und deutschen Elementen besteht und sich langsam, Schritt für Schritt, in einen vollständig übersetzten deutschen Satzbaum verwandelt.

Der deutsche Satzbaum wird dann schließlich noch durch das *Generierungsmodul* in die korrekte Oberflächenform gebracht. Hierzu muß ebenfalls der Satzbaum durchquert werden, um die einzelnen Worte in der richtigen Reihenfolge zu einem Satz zusammenzusetzen. Für die Bestimmung der korrekten Artikel und Flexionen wird sowohl auf die im Satzbaum angegebenen Eigenschaften zurückgegriffen (z.B. Zeitstufe, Komparation, Bestimmtheit, Zahl) als auch auf Informationen aus dem deutschen Lexikon (z.B. Geschlecht).

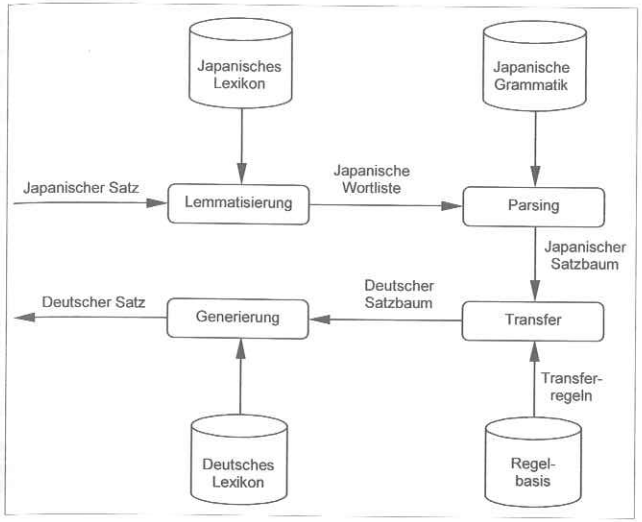

Abb. 6: Übersetzungsmodus

Das Übersetzungssystem wurde vollständig in Amzi! Prolog implementiert, eine logische Programmiersprache, die sich optimal für die deklarative Beschreibung von Grammatik- und Transferregeln eignet. Amzi! Prolog hat auch den entscheidenden Vorteil, daß es vollständige Unicode-Unterstützung bietet, d.h. man kann die japanischen Schriftzeichen ohne Probleme im Programmcode verwenden. Weiters werden komfortable Schnittstellen angeboten, um Amzi! Prolog in andere Programme einzubetten, insbesondere auch in Microsoft Word, was es erst ermöglicht, alle Übersetzungsfunktionen direkt aus dem Texteditor heraus aufzurufen, ohne überhaupt zu merken, daß Amzi! Prolog im Hintergrund läuft. Schließlich ist ein letztes wichtiges Argument die ausgezeichnete Skalierbarkeit von Amzi! Prolog, die garantiert, daß auch noch bei sehr großen Lexika, Grammatiken und Regelbasen die Übersetzungen ohne merkbare Verzögerung durchgeführt werden können.

Mein persönlicher Übersetzungs- und Leseassistent

Nach Fertigstellung des Übersetzungssystems war der nächste Schritt, dieses mit dem bereits zuvor entwickelten Leseassistenten zu integrieren, um Sprachstudenten eine umfassende Lernumgebung zur Verfügung zu stellen, welche von mir den Namen *PETRA (Personal Embedded Translation and Reading Assistant)* erhielt. PETRA kann direkt aus Microsoft Word für jedes beliebige Dokument an einer bestimmten Textstelle durch eine einfache

Tastenkombination aktiviert werden. Es wird für die aktuelle Cursorposition automatisch der diese umgebende japanische Satz ermittelt und in die PETRA-Arbeitsumgebung kopiert, die als eigenes Fenster dargestellt wird. Abbildung 7 zeigt ein Beispiel der Benutzeroberfläche für ein japanisches Dokument über die Entstehungsgeschichte des gedruckten Buchs. In diesem Beispiel hat ein Sprachstudent PETRA für den fünften Satz aktiviert, die Aussprache und Bedeutung des ersten Wortes im Satz im Wörterbuch nachgeschlagen und sich schließlich den Übersetzungsvorschlag von PETRA für den vollständigen Satz anzeigen lassen. Als zusätzliche wichtige Informationen kann der Student auch japanische und deutsche Wortlisten und Satzbäume abfragen, um auf diese Weise relevante Details über Wortarten sowie über syntaktische Eigenschaften und Strukturen zu erfahren.

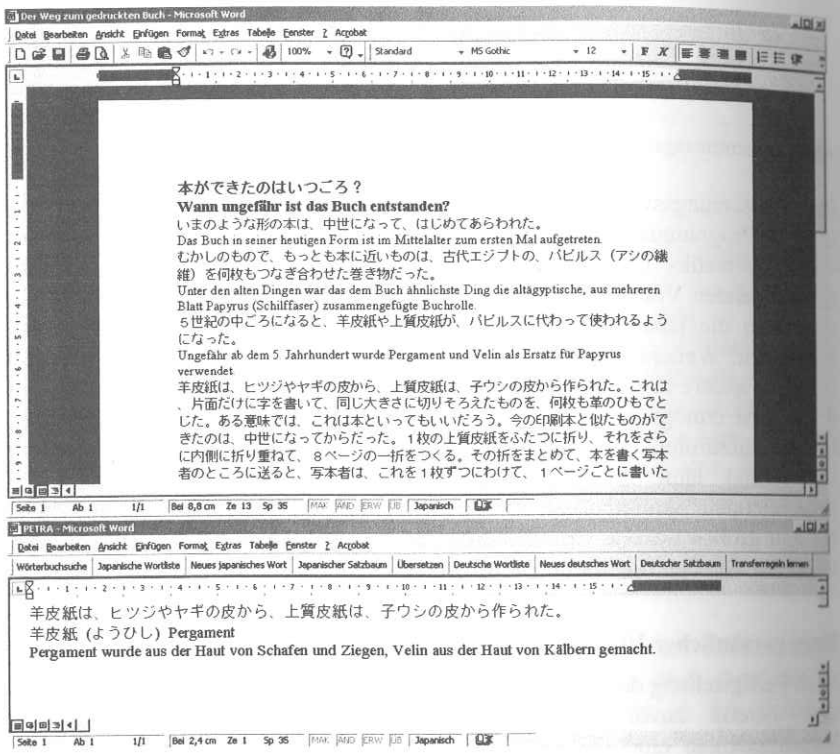

Abb. 7: Beispiel für Benutzeroberfläche

Schließlich habe ich noch benutzerfreundliche Formulare entwickelt, um neue Wörter unter Angabe aller notwendigen syntaktischen Eigenschaften zum japanischen oder deutschen Lexikon hinzufügen zu können, z.B. zeigt Abb. 8 das Formular für deutsche Nomen.

Abb. 8: Beispiel für Formular

Erste Erfahrungen mit PETRA und Ausblick

Nachdem ich die Implementierung von PETRA abgeschlossen hatte, habe ich mit einigen Freiwilligen erste Experimente für den praktischen Einsatz durchgeführt. Ganz allgemein war das Feedback der Studenten sehr positiv. Die Arbeit mit PETRA macht Spaß, vermittelt ein neues Verständnis der strukturellen Zusammenhänge und regt zum aktiven Lernen an. Besonders positiv hervorgehoben wurde der interaktive Charakter, der zum Experimentieren einlädt. Man kann einfach unterschiedliche Dinge ausprobieren und sofort sehen, wie sie sich auswirken.

Bezüglich der Oberflächengestaltung habe ich auch mit anderen Varianten experimentiert und die Reaktion der Studenten beobachtet. Die Verwendung von Pop-up-Fenstern, um kurzfristig Informationen einzublenden, hat sich nicht bewährt. Die überwiegende Rückmeldung war, daß solche Informationen, z.B. Wortbedeutungen, dauerhaft sichtbar sein sollen. Da Studenten bei komplexen Sätzen oft längere Zeit für die Interpretation benötigen, sich verschiedene Formulierungen überlegen, etc., wird es als sehr störend empfunden, wenn man immer wieder aufs Neue Pop-up-Fenster aktivieren muß. Ich habe auch

Versuche durchgeführt, bei denen die Studenten direkt im japanischen Dokument gearbeitet haben, d.h. ohne eigenes Fenster für PETRA. Auch hier waren die Erfahrungen eher negativ, da die Studenten es als unübersichtlich bewertet haben. Die Vermischung von Originaltext und Hilfestellungen führte eher zu Verwirrung und lenkte vom Textzusammenhang ab. Schließlich kristallisierte sich als ein wichtiges Designkriterium heraus, den Studenten nur soviel Information anzubieten, wie diese wirklich – ihrem jeweiligen Wissensstand entsprechend – benötigen. Aus diesem Grund muß z.B. der Student einzelne Worte im Wörterbuch nachschlagen anstatt daß ihm gleich für jedes Wort im Satz alle möglichen Bedeutungen angezeigt werden. Auch hier gilt, daß oft zuviel Information sich wiederum negativ auswirkt. Die Studenten haben es als lästig empfunden, wenn ihnen unnötige Hilfestellungen gegeben werden. Sie wollten lieber selbst aktiv die Information auswählen, die sie wirklich brauchten. Diese selektive Unterstützung hat auch einen wesentlichen pädagogischen Nutzen, denn wenn Studenten die Aufgabe zu leicht gemacht wird, reduziert sich auch der erzielte Lerneffekt.

Durch die Protokollierung der Herangehensweise der Studenten an Übersetzungsaufgaben mit PETRA konnten wir folgende „Best Practice" ableiten. Ein Student sollte immer zuerst versuchen, einen japanischen Satz korrekt zu segmentieren und die Bedeutungen der einzelnen Worte herauszufinden. Wenn notwendig, kann er hierfür das Wörterbuch und die Anzeige der Wortliste zu Hilfe nehmen. Danach sollte er sich über die Satzstruktur Klarheit verschaffen, wobei ihn die Anzeige des Satzbaums unterstützt. Schließlich soll ein eigener Übersetzungsvorschlag formuliert und erst dann dieser mit dem Übersetzungsvorschlag von PETRA verglichen werden. Auf der Grundlage der beiden Vorschläge sollte man dann zu einer zufriedenstellenden Lösung konvergieren. Dies kann erfordern, daß der Student seinen eigenen Vorschlag überdenkt oder aber das System dazu veranlaßt, seine Regelbasis zu überarbeiten. Es kommt daher zu einem bidirektionalen Wissensaustausch, indem einerseits der Student von PETRA, aber auch PETRA vom Studenten lernt. Somit übernimmt der Student eine aktive Rolle und kann gleichzeitig PETRA vollständig auf seine persönlichen Präferenzen abstimmen.

Zur Zeit überlege ich, die Möglichkeit in PETRA vorzusehen, ein solchermaßen strukturiertes Ablaufmodell vom System her vorzugeben, d.h. daß PETRA dann dem Studenten in stärkerem Maße Anleitung gibt und ihn Schritt für Schritt durch den Prozeß führt. In nächster Zeit ist weiters eine größer angelegte Evaluierung des Systems vorgesehen, sowohl in Hinblick auf Übersetzungsqualität als auch auf die Effektivität für den Einsatz im Sprachunterricht.

Quellenangaben

P. Brown: *A statistical approach to machine translation.* Computational Linguistics 16(2), 1990.

J. Hutchins: „Machine translation and computer-based translation tools: What's available and how it's used" In: J. M. Bravo, ed.: *A New Spectrum of Translation Studies*. University of Valladolid, 2003.

S. Sato: *Example-based Machine Translation.* Dissertation, Kyoto University, 1991.

H. Somers, ed.: *Computers and Translation: A Translator's Guide.* John Benjamins, 2003.

W. Winiwarter: *A language learning environment for assisting foreigners in reading Japanese Web pages. Proceedings of the 5th International Congress on Terminology and Knowledge Engineering,* Innsbruck, Österreich, 1999.

W. Winiwarter: *Incremental learning of transfer rules for customized machine translation. Proceedings of the 15th International Conference on Applications of Declarative Programming and Knowledge Management,* Berlin, Deutschland, 2004.

Links

Amzi Prolog!: http://www.amzi.com
ATLAS: http://www.translation.net/atlas.html
EDICT: http://www.csse.monash.edu.au/~jwb/wwwjdic.html
POPjisho: http://www.popjisyo.com
Rikai: http://www.rikai.com/perl/Home.pl
TRADOS: http://www.trados.com
WaDokuJT: http://www.wadoku.de

Weiterführende Literatur

J. Hutchins: *Machine Translation: Past, Present, Future.* Ellis Horwood, 1986.

J. Hutchins: *Machine translation over 50 years. Histoire epistémologie langage 23(1)*, 2001.

J. Hutchins: *Has machine translation improved? Some historical comparisons. Proceedings of the 9th MT Summit*, New Orleans, USA, 2003.

J. Hutchins, H. Somers: *An Introduction to Machine Translation.* Academic Press, 1992.

J. Newton, ed.: *Computers in Translation: A Practical Appraisal.* Routledge, 1992.

Ao. Univ.-Prof. Mag. Mag. Dr. Werner Winiwarter (* 1966). 1985-1991 Studium der Wirtschaftsinformatik sowie der Politikwissenschaft und Japanologie an der Universität Wien. 1995 Promotion an der Universität Wien im Bereich Sprachtechnologie. Herbst 1995 bis Herbst 1997 Forschungsaufenthalt am Department of Information Science an der Kyoto University, Erwin-Schrödinger-Auslandsstipendium, Thema der Forschungsarbeit: *Frageunterstützung für das Unterrichtssystem VIEW Classroom.* April 1998 Habilitation an der Universität Wien für das Fach Angewandte Informatik, Thema der Habilitationsschrift: *Natural Language Engineering.* Frühjahr 2000 – Herbst 2001 Gastprofessur an der Johannes Kepler Universität Linz und Wissenschaftlicher Leiter des Software Competence Center Hagenberg. Herbst 2001 bis Frühjahr 2002 Wissenschaftlicher Geschäftsführer des Electronic Commerce Competence Center in Wien. Seit 1997 regelmäßige kurzfristige Forschungsaufenthalte an der Kyoto University. Forschungsschwerpunkt Sprachtechnologie, speziell seit 1997 maschinelle Übersetzung und computergestützter Spracherwerb für Japanisch. Weitere Schwerpunkte: eBusiness, Semantic Web, maschinelle Lernverfahren, Unterrichtssysteme, Information Retrieval und Sicherheit. Über 100 internationale Publikationen. Organisation und Mitgliedschaft in Programmkomitees von zahlreichen internationalen Konferenzen und Workshops.