Sehen und Verstehen: Der Beitrag bildlicher Information zur robusten Sprachverarbeitung

Wolfgang Menzel

Fachbereich Informatik, Universität Hamburg
Vogt-Kölln-Straße 30, 22527 Hamburg, Germany

Zusammenfassung Sprachverstehen ist in hohem Maße situativ beeinflußt. Um diese Charakteristik des menschlichen Vorbilds in technische Lösungen zu übertragen, wird eine Verarbeitungsarchitektur entwickelt, die in der Lage ist, sehr unterschiedliche informationelle Beiträge in einem einheitlichen Entscheidungsverfahren zusammenzufassen. Am Beispiel eines Einsatzes in Lehrsystemen für den Fremdsprachenunterricht wird gezeigt, wie propositional repräsentierte piktorielle Information eine robuste Analyse von möglicherweise hochgradig gestörten sprachlichen Äußerungen unterstützen kann.

1 Einführung

Technische Lösungen zur Integration von Bild- und Sprachverstehen beruhen üblicherweise auf der relativ losen Kopplung von zwei weitgehend eigenständigen Komponenten (vgl. Bild 1a). Da eine solche Anordnung keinerlei direkten Informationsaustausch zwischen den parallelen Verarbeitungssträngen erlaubt, erfolgt die Analyse von Bild- und Sprachdaten praktisch unabhängig von dem aktuellen Informationsstand auf der jeweils komplementären Wahrnehmungsebene und die Resultate dieser Berechnungsprozesse werden erst auf einer relativ tiefen Verarbeitungsstufe miteinander in Beziehung gesetzt. Demgegenüber zeichnet sich die Kopplung der Sinnesmodalitäten beim menschlichen Vorbild durch ein erheblich stärkeres Zusammenwirken aus.

Bereits auf der rein phonetischen Ebene läßt sich eine sehr starke Beeinflussung der sprachlichen Perzeption durch die visuelle Beobachtung des Sprechvorgangs nachweisen. Einerseits führen etwa desynchronisierte Videoaufzeichnungen der Lippenbewegungen systematisch zu fehlerhaften Lautwahrnehmungen [9], während andererseits unter Rückgriff auf den Informationsbeitrag der Lippenbewegungen auch im Fall eines extrem gestörten akustischen Kanals (verminderte Hörfähigkeit, starke Umgebungsgeräusche usw.) eine Verständigung noch ermöglicht werden kann.

Durch eine erhebliche Verfeinerung der Techniken zur Verfolgung von Augenbewegungen konnte eine enge Kopplung von Sprach- und Bildverarbeitung auch für die höheren Ebenen der Sprachverarbeitung nachgewiesen werden [22]. So erfolgt die visuelle Identifizierung der sprachlich vermittelten Objektreferenzen auch im Falle komplexer sprachlicher Konstruktionen unmittelbar, sobald ausreichend viel Information verfügbar ist. Zudem läßt sich zeigen, daß die durch den visuellen Kontext bereitgestellte Information die syntaktische Verarbeitung in einer Weise beeinflußt, wie dies für sprachliche Kontexte bereits von Steedman und Crain [1] behauptet worden war. Insbesondere folgt die Aufmerksamkeitssteuerung in den Fällen eindeutiger Referenzbedingungen nicht den durch das Sprachsystem vorgegebenen Präferenzen, was mit der Annahme einer weitgehend autonomen syntaktischen Verarbeitung im sprachlichen Verarbeitungskanal [5] nicht verträglich ist.

Da vermutet werden kann, daß die enge Kopplung des Sprachverstehens mit der visuellen Wahrnehmung auch auf den höheren Ebenen der Sprachverarbeitung einen erheblichen Beitrag zur Robustheit des Gesamtsystems erbringt [11], stellt sich die Frage nach geeigneten Systemarchitekturen, die ein derartiges Zusammenwirken bereits zu einem relativ frühen Zeitpunkt überhaupt ermöglichen. Ein spezieller Bedarf an derartigen Lösungen ergibt sich vor allem im Bereich der Verarbeitung gesprochener Sprache, sowie bei der Analyse fehlerhafter sprachlicher Konstruktionen, die etwa im Kontext des Fremdsprachenunterrichts besonders häufig auftreten.

Aufbauend auf einer Analyse existierender Lösungsansätze aus diesen beiden Anwendungsfeldern werden allgemeine Anforderungen an eine geeignete Systemarchitektur identifiziert und in einen Lösungsvorschlag umgesetzt, der auf einem Verfahren zum eliminativen Parsing in Mehrebenenrepräsentationen beruht. Eine erste experimentelle Umsetzung erfolgt in einem Lehrsystem für den Fremdsprachenunterricht.

2 Sprachsignalerkennung

Tatsächlich finden sich erste Lösungsansätze zur wechselseitigen Beeinflussung von Bild- und Sprachverarbeitung vorrangig in Szenarien zur (gesprochen-) sprachlichen Interaktion mit autonomen Robotern, wo von dem Zusammenwirken u.a. eine Reduktion der Erkennungsunsicherheit in beiden Wahrnehmungskanälen erwartet wird. Naeve u.a. [17] nutzen die aus dem visuellen Kanal resultierenden Angaben über mögliche Referenzobjekte, um spezielle, der jeweiligen Situation angepaßte Sprachmodelle (in Form von endlichen Automaten) zu generieren. Auf diese Weise gelingt es, die Gesamtperplexität fast um einen Faktor zwei zu reduzieren und die entsprechenden visuell induzierten Erwartungen auf die Ebene der Sprachsignalerkennung abzubilden. Die Kopplung der visuellen und sprachlichen Information erfolgt auf der Ebene der Domänenkonzepte über die eine bidirektionale Ausbreitung von Informationsbeiträgen aus den beiden Wahrnehmungskanälen erfolgen kann.

Handelt es sich bei diesem Verfahren um einen Lösungsansatz zur Steigerung der Erkennungssicherheit für Objektbenennungen, wendet sich Socher [21] dem Problem der Objektidentifizierung zu. Bayes'sche Netze werden benutzt, um visuelle und sprachliche Informationsbeiträge zur Referenzidentifikation zusammenzuführen. Wegen der bidirektionalen Arbeitsweise der Netze, die den Transport kausaler und diagnostischer Evidenz unterstützt, ist auch hier eine Kopplung der beiden Modalitäten über die Konzeptebene möglich.

Für den jeweils ausgewählten Teilbereich liegt beiden Ansätzen eine Anordnung der Verarbeitungskomponenten zugrunde, die über die gerichtete In-

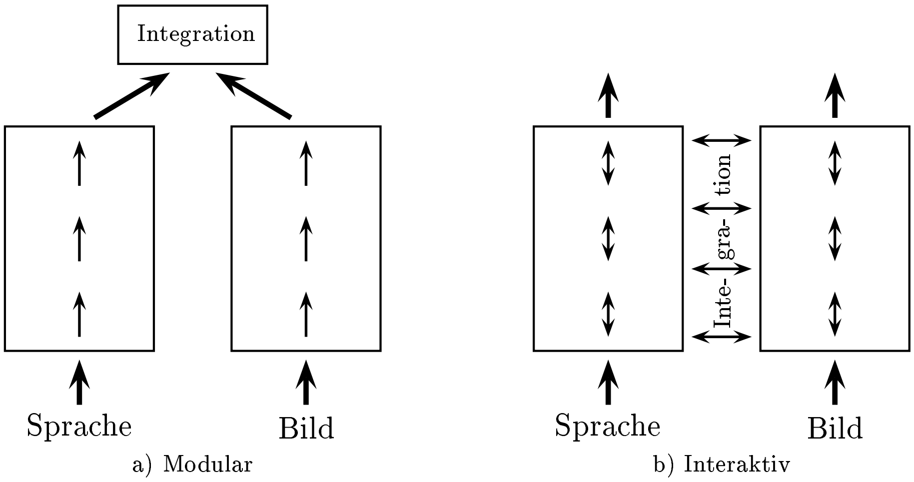

Abbildung1. Systemarchitekturen für die Integration komplementärer Informationsbeiträge

formationausbreitung in Bild 1a hinausgeht. Damit bieten sie tatsächlich die notwendigen Voraussetzungen für die gewünschte Integration von visueller und sprachlicher Information zumindest für isolierte Phänomenbereiche. Basis dafür ist jedoch eine allein auf Hypothesenprädiktion abzielende, strikt generative Sichtweise. Gegen deren Erweiterbarkeit beim Übergang von einfachen Prädikationen auf komplexe relationale Zusammenhänge oder gar temporale Abläufe spricht vor allem die Tatsache, daß die rein generative Sicht

- bei größeren Domänen sehr schnell auf prinzipielle Realisierungsschwierigkeiten stößt und
- die Perspektive des Systementwicklers sehr starr[1] auf die sprachliche Realisierung abbildet.

Komplementär hierzu sind die eher analytisch ausgerichteten Ansätze angelegt, die sich an den klassischen Parsingverfahren orientieren, allerdings kaum auf die Erfordernisse einer Interaktion mit externen Wissensquellen Rücksicht nehmen. Beispiele hierfür sind die Systemrealisierungen in [16] und [2], denen aber wiederum eine Systemarchitektur nach Bild 1a zugrunde liegt. Die Kopplung der beiden unidirektionalen Wahrnehmungskanäle erfolgt erst nach dem Parsing, wodurch eine direkte Beeinflussung der Sprachverarbeitung durch die visuellen Befunde ausgeschlossen ist.

3 Sprachlehrsysteme

Recht ähnlich stellt sich die Situation im Bereich der Lehrsysteme für den Fremdsprachenunterricht dar. Hier besteht bereits aus didaktischen Überlegungen heraus die Notwendigkeit, sprachliche Interaktion in einen Kontext einzubetten, um einerseits die Aufgabenstellung für den Schüler zu motivieren und andererseits

[1] Mit Ausnahme der Möglichkeit zum Ignorieren von Signalabschnitten.

einen kommunikativen Zusammenhang herzustellen. Besonders gut eignen sich hierfür direkt manipulierbare virtuelle Welten, in denen Schüler und Lehrsystem gleichzeitig sprachlich und grafisch agieren. Wiederum ist eine Integration und Abstimmung der verschiedenen Informationsbeiträge gefordert. Insbesondere besteht aber die Hoffnung, daß die vom Lehrsystem ohnehin verwaltete grafische Information zur Unterstützung der Analyse und der Diagnose der oftmals hochgradig fehlerhaften Spracheingabe herangezogen werden kann.

Klassische unidirektionale Parser (z.B. [4], aber auch [6, 19, 20, 23, 25]) orientieren sich ebenfalls an einer modularen Architektur nach Bild 1a. Um die erforderliche Robustheit gegenüber Schülerfehlern zu erreichen, kommen unterschiedliche Varianten der Idee des *minimal distance parsing* [18] zum Einsatz. Dabei wird die zur Fehlererklärung herangezogene Diagnose aus derjenigen validen Strukturbeschreibung abgeleitet, die eine möglichst geringe Distanz zu der vorgefundenen Schülerlösung aufweist. Das Abstandsmaß ist dabei über Anzahl und Art der zur Lösungsfindung ggf. erforderlichen Fehlerregeln bzw. Bedingungsrücknahmen definiert. Da beide Techniken die zur Suchraumbegrenzung zwingend erforderlichen Einschränkungen neutralisieren, ist eine Generalisierbarkeit problematisch, solange hierfür keine adäquate Kompensation z.B. durch Restriktionen aus dem visuellen Kontext erfolgt.

Eine (rudimentäre) Steuerung durch den (sprachlichen) Kontext ist nur über die Zuweisung von Straffaktoren zu Grammatikregeln möglich. Da Grammatikregeln aber kaum direkte Korrelationen mit den außersprachlichen Gegebenheiten des visuellen Kontexts aufweisen, ist dieser Mechanismus für die Integration visueller Information kaum geeignet.

Wesentlich bessere Voraussetzungen für die Einbeziehung außersprachlicher Bedingungen ergeben sich in einem erwartungsgesteuerten Ansatz, wie er beispielsweise in [3] realisiert wurde. Ausgehend von einem durch das finite Verb des Satzes instanziierten Prädikat werden die Argumentstellen durch das sprachliche Material gefüllt, wobei wiederum ein geeignet zu definierendes Ähnlichkeitsmaß für die Auswahl einer bestimmten Zuordnung maßgeblich ist (*predication-driven parsing*). Die Integration visuell präsentierter Informationen in diesen Auswahlprozeß erscheint prinzipiell möglich, wenn es gelänge, den zugrundeliegenden Abbildungsmechanismus auf mehrere alternative Prädikate zu erweitern und bei der Selektion der zu erwartenden Prädikate auch außersprachliche Information zu berücksichtigen.

4 Anforderungen an integrative Architekturen

Ausgehend von den bisherigen Beobachtungen in den genannten Anwendungsbereichen können eine Reihe von Anforderungen an eine Architektur zur integrativen multimodalen Verarbeitung formuliert werden:

Mehrebenenrepräsentation: Die Architektur muß die Verwaltung mehrerer zueinander komplementärer (d.h. parallel aufzubauender) Repräsentationsebenen unterstützen, zwischen denen ein Informationsaustausch erfolgen kann.

Feinkörnige Interaktion: Die wechselseitige Bezugnahme zwischen den beteiligten Repräsentationen muß auf allen Ebenen der hierarchischen Dekomposition erfolgen. Es reicht hierbei nicht aus, komplette Analyseergebnisse des visuellen und sprachlichen Kanals zu kombinieren, vielmehr muß auch für Teilergebnisse eine unmittelbare Wechselwirkung möglich sein.

Omnidirektionale Informationsausbreitung: Unterstützende Information muß zwischen allen Repräsentationsebenen austauschbar sein, ohne daß ein explizites Umschalten zwischen verschiedenen Ausbreitungsrichtungen erforderlich ist. Dies ist insbesondere deshalb notwendig, weil die Qualität der jeweils zur Verfügung stehenden Information (z.B. sichere vs. unsichere, aber auch relevante vs. irrelevante Information) nicht a priori bewertet werden kann, sondern sich üblicherweise erst im Verlaufe der Integration ergibt.

Behandlung widersprüchlicher Aussagen: Sowohl innerhalb einer Repräsentationsebene als auch ebenenübergreifend müssen Inkonsistenzen zwischen Informationsbeiträgen toleriert werden, um einerseits der inhärenten Unsicherheit der Kategorienwahrnehmung Rechnung zu tragen, aber auch um potentielle Widersprüche zwischen visueller Realität und sprachlicher Äußerung detektieren und diagnostizieren zu können.

Damit ergibt sich eine Systemarchitektur nach Bild 1b. Zwei parallele Verarbeitungskomponenten mit jeweils separaten Repräsentationen für bildliche und sprachliche Information sind über Interaktionspfade so miteinander gekoppelt, daß die entsprechenden Strukturen auf allen jeweils verfügbaren Repräsentationsebenen unmittelbar aufeinander abgebildet werden können. So sollte das Vorliegen einer bestimmten räumlichen Relation (z.B. ein Buch liegt auf dem Tisch) bestimmte präpositionale Konstruktionen (z.B. *"Buch auf dem Tisch"*) im sprachlichen Kanal präferieren und umgekehrt. Gleichzeitig muß aber vermieden werden, den Informationsbeitrag einer Verarbeitungskomponente der jeweils komplementären Repräsentation aufzuzwingen, weil dann Widersprüche (z.B. das Buch liegt in Wirklichkeit gar nicht auf dem Tisch, sondern im Regal) nicht mehr kommunikativ behandelbar sind. Zudem kann ja niemals mit Sicherheit davon ausgegangen werden, daß sich der Sprecher auf die Verwendung der visuell erfaßbaren Objekte beschränkt. In solchen Fällen ist die visuell verfügbare Information irrelevant für die Verarbeitung der sprachlichen Konstruktion und darf den Sprachanalyseprozeß nicht negativ beeinflussen.

5 Informationsintegration im eliminativen Parsing

Auf der Grundlage eines Verfahrens zum eliminativen Parsing wurde eine Mehrebenenarchitektur zur Integration von komplementären Wissensquellen realisiert [12, 8, 14]. Um die wechselseitige Abbildung der Repräsentationsstrukturen zu unterstützen, wird auf allen Ebenen ein einheitliches, hierarchisch strukturiertes relationales Darstellungsformat verwendet, das sich an dem Formalismus einer Dependenzgrammatik orientiert.

Modelliert wurden die Ebenen Syntax, Semantik und Domänenwissen (Bild 2). Lokale Konfigurationen aus maximal zwei Dependenzrelationen werden durch gewichtete Constraints lizensiert, wobei die Gewichte den Grad widerspiegeln, zu dem eine Verletzung eines Constraints noch akzeptabel erscheint. Auf diese Weise können sehr unterschiedliche sprachliche und außersprachliche Wissensbestandteile modelliert werden:

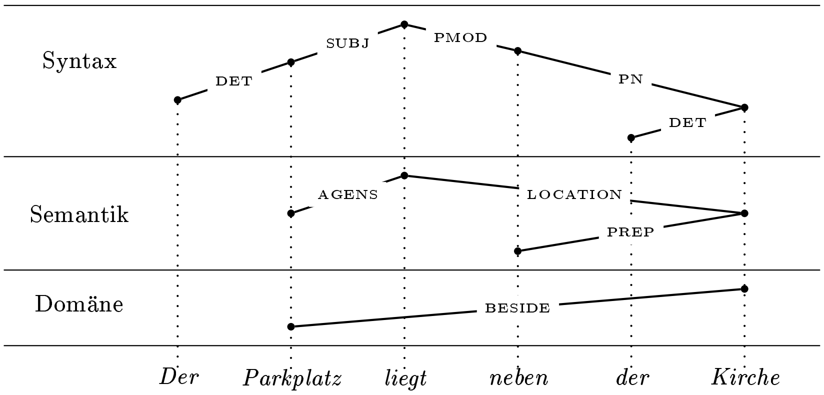

Abbildung2. Mehrebenenrepräsentation für eine Schüleräußerung.

Harte Constraints müssen von jeder Strukturbeschreibung erfüllt werden. Sie dienen in erster Linie dazu, Wohlgeformtheitsbedingungen für Strukturbeschreibungen zu formulieren (z.B. Relationen wie SUBJECT-OF oder OBJECT-OF modifizieren immer ein Verb).

Defaults sind mögliche, aber hart bestrafte Strukturbeschreibungen, die nur dann "überleben", wenn im Hypothesenraum keine Alternative mehr verbleibt (z.B. kann bei Bedarf auch ein Nomen als Spitzenknoten einer nichtsatzwertigen Äußerung verwendet werden). Auf diese Weise wird die Möglichkeit zum partiellen Parsing problemlos in den Ansatz integriert.

Korrektheitsbedingungen stellen die Wohlgeformtheit der sprachlichen Ausdrücke auf den verschiedenen Ebenen sicher (z.B. Obligatheitsforderungen, Kongruenz- und Rektionsbedingungen, selektionale Restriktionen, Domänenwissen usw.). Die Bewertung variiert in Abhängigkeit von der Verläßlichkeit der jeweiligen Forderungen.

Präferenzen differenzieren zwischen sprachlichen Optionen innerhalb des Bereiches der Wohlgeformtheit (z.B. Subjekt-vor-Objekt-Präferenz).

Beginnend mit einer vollständig unterspezifizierten Ausgangsstruktur, die grundsätzlich alle durch das Relationeninventar definierten Dependenzrelationen in einer kompakten Form enthält, wird durch die Anwendung von Constraints der Raum möglicher Lösungen sukzessive eingeschränkt, bis nur noch eine eindeutige Strukturbeschreibung auf den verschiedenen Repräsentationsebenen vorliegt. Dabei erlaubt die parallele Verfügbarkeit aller Strukturbeschreibungen eine vergleichende Bewertung für lokale Strukturkonfigurationen. In Abhängigkeit von den Ergebnissen der Constraintanwendung werden diejenigen Dependenzrelationen gelöscht, für die ein Maximum an negativer Evidenz akkumuliert wurde.

Da auch die bidirektionale Abbildung zwischen Dependenzrelationen auf den verschiedenen Repräsentationsebenen durch gewichtete Constraints erfolgt, kann so der gewünschte ebenenübergreifende Abgleich von Evidenz erfolgen. Steht die ermittelte Strukturbeschreibung im Widerspruch zu einer oder mehreren Korrektheitsbedingungen, so können diese Constraintverletzungen auch als Diagnose

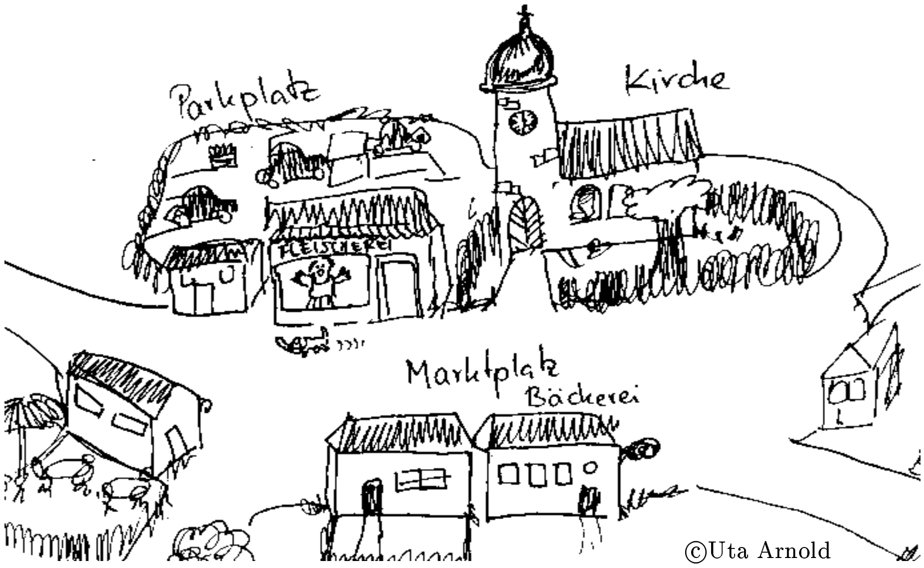

Abbildung3. Ausschnitt aus einem Kleinstadt-Szenario.

der Schüleräußerung interpretiert und in geeignete Fehlererklärungen umgewandelt werden [15]. Dieser enge Bezug zwischen den Parsingresultaten und der Diagnosefähigkeit bildet die Grundlage für eine Integration des Verfahrens in ein prototypisches Sprachlehrsystem [13]. Dazu wurde die Domänenebene durch das propositional repräsentierte Wissen aus der bildlichen Darstellung einer Szenerie modelliert (Bild 3), wobei sich der Prototyp durch eine Reihe sehr gravierender Vereinfachungen auszeichnet:

- Es handelt sich um eine statische Welt.
- Es wird von einer einfachen Referenzauflösung ausgegangen, d.h. nominale Bezeichner stehen für die Objektinstanzen.
- Es wird ausschließlich eine extrinsische Perspektive zugrundegelegt.
- Es werden ausschließlich relativ einfache, narrative Sprachäußerungen betrachtet.

Tatsächlich stellt sich bei geeigneter Modellierung sehr schnell das angestrebte kompensatorische Verhalten ein, wobei insbesondere für fehlerhafte Äußerungen eine wechselseitige Unterstützung der Repräsentationsebenen zu beobachten ist. Im Normalfall wird semantische und domänenspezifische Information benutzt, um die syntaktische Diagnose zu steuern, etwa zur Subjekt-Objekt-Disambiguierung in Fällen wie

**Dort entdeckt Verkäuferin ein kleines Hund.*

oder aber zur Klärung der PP-Anbindung in

**Mich gefällt Auto auf das Parkplatz.*

Wegen des bidirektionalen Charakters der ebenenübergreifenden Constraints entsteht hier allerdings keine einseitige (und damit in einigen Fällen fatale) Abhängigkeit von einer bestimmten Ebene, wie sie etwa für semantisch gesteuerte Ansätze durchaus typisch ist. Solange ausreichend stabile syntaktische Kriterien vorhanden sind, stützen diese die Disambiguierung auf den komplementären Ebenen, so daß bei (relativ zum Domänenwissen) unbekannten Sachverhalten in Fällen wie

Der Parkplatz liegt neben dem Krankenhaus.

eine Extraktion des (noch unbekannten) propositionalen Gehalts versucht werden kann. Darüberhinaus bietet das Verfahren aber auch bei (relativ zum Domänenwissen) sachlich fehlerhaften Beschreibungen der Art

Die Bäckerei liegt hinter dem Parkplatz.

eine Diagnose der vorliegenden Inkonsistenzen an, die dann im Dialog mit dem Schüler weiter thematisiert werden kann. Das Verfahren realisiert somit eine durch den visuellen Kontext gesteuerte Sprachanalyse, ohne jedoch das Analyseergebnis durch die Spezifikation der Kontextinformation vollständig zu determinieren, so daß ein Verlassen des vorgegebenen Szenarios unter bestimmten Bedingungen (u.a. ausreichende lexikalische Abdeckung) immer noch möglich ist.

6 Bildliche Information in Sprachlehrsystemen

Mit dem Vorliegen eines Verfahrens, das einen bidirektionalen Informationsaustausch zwischen visueller und sprachlicher Information vermittelt, sind die Voraussetzungen für eine ganze Reihe von didaktisch interessanten Interaktionsformen im Kontext eines Sprachlehrsystems gegeben. Hamburger [7] unterscheidet acht Interaktionstypen, für die aus der hier betrachteten Sprachanalysesicht vor allem die folgenden fünf von Bedeutung sind:

Quizmaster: Das System fragt und der Schüler antwortet.

Movecaster: Das System generiert eine (animierte) Grafik und der Schüler erzählt, was er sieht.

In beiden Fällen überprüft das System die Akzeptabilität der Schüleräußerung in sprachlicher und pragmatischer Hinsicht, wobei die grafisch präsentierte Information als Stütze bei der Behandlung fehlerhafter Konstruktionen dient. Als Interaktionsformen, bei denen die Initiative überwiegend beim Schüler liegt, kommen in Frage:

Oracle: Der Schüler fragt und das System antwortet.

Servant: Der Schüler gibt Kommandos und das System führt sie aus.

Interpreter: Der Schüler erzählt etwas und das System generiert eine entsprechende Animation.

Darüberhinaus kann die interaktive Manipulation einer grafischen Repräsentation aber auch als alternativer Kommunikationskanal benutzt werden, falls wegen der eventuell noch mangelhaften Sprachbeherrschung eine sprachliche Verständigung nicht mehr möglich ist. Indem der Schüler seine sprachliche Äußerung gleichzeitig direktmanipulativ visualisiert, enthält das System zusätzliche Informationen, die zur Steuerung der Diagnose verwendet werden können. Mealing und Yazdani [10, 24] schlagen für diesen Zweck eine kompositionelle, und teilweise animierte Piktogramm-Sprache vor, die eine eindeutige Visualisierung sprachlicher Äußerungen gestattet.

7 Schlußfolgerungen

Unter Verwendung von Techniken des eliminativen Parsings konnte eine Systemarchitektur realisiert werden, die die Forderung nach Mehrebenenrepräsentation, feinkörniger Interaktion, omnidirektionalem Informationsfluß und Behandlung widersprüchlicher Information erfüllt. Sie gestattet die Beinflussung der Sprachverarbeitung durch visuell gegebene Information, ohne daß dadurch eine einseitige Dominanz des visuellen Kanals in Kauf genommen werden muß. Wesentlicher Vorteil dieses integrativen Ansatzes ist seine Robustheit gegenüber fehlerhaftem sprachlichen Input, verbunden mit der Fähigkeit zur Diagnose von Inkonsistenzen auf allen Repräsentationsebenen. Dies erlaubt die Entwicklung von Lehrsystemen für den Fremdsprachenunterricht, die wegen der situativen Verankerung der Sprachverwendung eine deutlich gesteigerte kommunikative Relevanz der Schülerübungen vorweisen können.

Eindeutig unbefriedigend ist zur Zeit noch die Ausdruckskraft der visuellen Repräsentationsebene. Hier wird eine sukzessive Erweiterung der Möglichkeiten im Hinblick auf die Behandlung komplexerer Szenarien erforderlich sein, wobei jedoch darauf zu achten ist, daß die vorteilhaften Eigenschaften der Integrierbarkeit erhalten bleiben.

Literatur

1. S. Crain and M. Steedman. On not being led up the garden path. In D. R. Dowty, L. Kartunen, and A. M. Zwicky, editors, *Natural Language Parsing: Psychological, computational, and theoretical perspectives*, pages 320–358. Cambridge University Press, Cambridge, 1985.
2. C. Crangle and P. Suppes. *Language and Learning for Robots*, volume 41 of *CSLI Lecture Notes*. CSLI, Stanford, CA, 1994.
3. W. H. DeSmedt. Herr Kommissar: An ICALL conversation simulator for intermediate german. In V. M. Holland, J. D. Kaplan, and M. R. Sams, editors, *Intelligent Language Tutors: Theory Shaping Technology*, chapter 9, pages 153–174. Lawrence Erlbaum Ass., Mahwah, NJ, 1995.
4. S. Felshin. The Athena language learning project NLP system: A multilingual system for conversation-based language learning. In V. M. Holland, J. D. Kaplan, and M. R. Sams, editors, *Intelligent Language Tutors*, chapter 14, pages 257–272. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.
5. L. Frazier. Theories of sentence processing. In J. L. Garfield, editor, *Modularity in Knowledge Representation and Natural Language Understanding*, pages 291–307. MIT-Press, Cambridge MA, 1987.
6. L. K. Hagen. Unification-based parsing applications for intelligent foreign language tutoring systems. *CALICO Journal*, 12(2/3):5–30, 1995.

7. H. Hamburger. Tutorial tools for language learning by two-medium dialogue. In V. M. Holland, J. D. Kaplan, and M. R. Sams, editors, *Intelligent Language Tutors*, chapter 10, pages 183–199. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.
8. J. Heinecke, J. Kunze, W. Menzel, and I. Schröder. Eliminative parsing with graded constraints. In *Proceedings 17th International Conference on Computational Linguistics, 36th Annual Meeting of the ACL, Coling-ACL '98*, Montreal, Canada, 1998.
9. H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.
10. S. Mealing and M. Yazdani. A computer-based iconic language. In M. Yazdani, editor, *Multilingual Multimedia: Bridging the Language Barrier with Intelligent Systems*. Intellect Books, Oxford, 1993.
11. W. Menzel. Robust parsing of natural language. In *KI-95: Advances in Artificial Intelligence*, pages 19–34, Berlin, 1995. Springer-Verlag.
12. W. Menzel. Constraint satisfaction for robust parsing of natural language. *Theoretical and Experimental Artificial Intelligence*, 10(1):77–89, 1998.
13. W. Menzel and I. Schröder. Constraint-based diagnosis for intelligent language tutoring systems. In *Proceedings IT & KNOWS, XV. IFIP World Computer Congress*, Wien und Budapest, 1998.
14. W. Menzel and I. Schröder. Decision procedures for dependency parsing using graded constraints. In S. Kahane and A. Polguere, editors, *Proc. Coling-ACL Workshop on Processing of Dependency-based Grammars*, Montreal, Canada, 1998.
15. W. Menzel and I. Schröder. Model-based diagnosis under structural uncertainty. In *Proceedings 13th European Conference on Artificial Intelligence*, pages 284–288, Brighton, 1998.
16. J.-T. Milde, K. Peters, and S. Strippgen. Situated communication with robots. In *Proc. 1st. Int. Workshop on Human–Computer Conversation*, Bellagio, Italy, 1997.
17. U. Naeve, G. Socher, G. A. Fink, F. Kummert, and G. Sagerer. Generation of language models using the results of image analysis. In *4th European Confence on Speech Communication and Technology*, pages 1739–1742, Madrid, 1995.
18. C. P. Rosé and A. Lavie. An efficient distribution of labor in a two stage robust interpretation process. In *Proceedings EMNLP-2*, 1997.
19. C. B. Schwind. An intelligent language tutoring system. *International Journal of Man-Machine Studies*, 33:557–579, 1990.
20. C. B. Schwind. Error analysis and explanation in knowledge based language tutoring. *Computer Assisted Language Learning*, 8(4):295–324, 1995.
21. G. Socher. *Qualitative Scene Descriptions from Images for Integrated Speech and Image Understanding*, volume 170 of *Dissertationen zur Künstlichen Intelligenz*. Infix, Sankt Augustin, 1997.
22. M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, 1995.
23. A. Weinberger, J. Garman, J. Martin, and P. Merlo. A principle-based parser for foreign language tutoring in german and arabic. In V. M. Holland, J. D. Kaplan, and M. R. Sams, editors, *Intelligent Language Tutors*, chapter 2, pages 23–44. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.
24. M. Yazdani and S. Mealing. Communicating through pictures. Technical report, Department of Computer Science, Exeter, 1993.
25. M. Yazdani and J. Uren. Generalising language tutoring systems. In M. Yazdani, editor, *Multilingual Multimedia: Bridging the Language Barrier with Intelligent Systems*. Intellect Books, Oxford, 1993.