BMFT

gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil

Verbundvorhaben

Definitheit und Numerus. Anforderungen an den Transfer Japanisch – Englisch

Melanie Siegel

Universität Bielefeld

Vm

Memo 56
Dezember 1994

Dezember 1994

Melanie Siegel

Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschaft
– Computerlinguistik –
Pf 100 131
D 33 502 Bielefeld

Tel.: (0521) 106 - 2996
Fax: (0521) 106 - 2996
e-mail: MELANIE@COLI.UNI-BIELEFELD.DE

Gehört zum Antragsabschnitt: 12 Transfer(Japanisch)

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 101 G gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Inhaltsverzeichnis

1 Einleitung

Ein Problem des Transfers in der maschinellen Übersetzung von Japanisch nach Englisch ist fehlende Information über Numerus und Definitheit im Japanischen, die für die Wahl der englischen Artikel und die Nomenmarkierung gebraucht wird. Obwohl dieses Problem signifikant ist, beschäftigt sich die Forschungsliteratur kaum damit. Ein Ansatz aus der Literatur ist der von Murata/Nagao (1994), der heuristische Verfahren verwendet.

Wir basieren unsere Untersuchungen auf experimentell erhobenen Daten aus einem Experiment über deutsch-japanische gedolmetschte Terminaushandlungsdialoge (Siegel et al. 1993, Siegel 1993). Auf diese Weise können Phänomene bestimmt werden, die für die Domäne von VERBMOBIL relevant sind. Wir sehen unser Vorgehen in Übereinstimmung mit dem 'Sublanguage'-Ansatz (Kittredge 1987).

2 Das Problem

Japanische Nominalphrasen enthalten im Normalfall keine Informationen über Numerus und Definitheit, da es kaum Artikel und Numerusmarkierungen gibt. Diese

Information wird jedoch für die Generierung der englischen Nominalphrasen gebraucht und muß daher in verschiedenen Wissensquellen gesucht werden. Folgendes Beispiel verdeutlicht diesen Mismatch:

```
[II-1, 20-22][1]
J: e. kayōbi   wa   watashidomo no   tokoro de  wa  kyūjitsu
      Dienstag WA   wir         NO   Seite  DE  WA  Feiertag
   na no de. anō . tabun .  kaigi    ni   sanka suru koto
   COP             vielleicht Treffen NI   teilnehmen NOM
   wa  dekimasen
   WA  können NEG
(On our side Tuesday is a holiday, maybe (we) cannot attend
the meeting.)
```

Die Informationen, daß „*kyujitsu*“ singular und indefinit ist und daher „*holiday*“ im Englischen den Artikel „*a*“ bekommen muß und daß „*kaigi*“ singular und definit ist und daher „*meeting*“ den Artikel „*the*“ bekommen muß, wird nicht vom Parser als Analyseergebnis geliefert. Lediglich in einigen wenigen Fällen ist eine Numerusmarkierung an Nomen mit Bezug auf Personen vorhanden. Im Beispiel ist es das Suffix „*domo*“ an „*watashidomo*“. Eine andere Möglichkeit dieser Art ist „ *gata*“ in „*katagata*“ (Personen).

Für die maschinelle Übersetzung, die sich mit der japanischen Sprache beschäftigt, ist dieses Problem von großer Relevanz, wie leicht einsichtig ist, da es bei jedem Auftreten einer Nominalphrase gelöst werden muß. Der Parser liefert eine Analyse ohne Informationen über Numerus und Definitheit, so daß hier die semantische Auswertung aktiv werden muß.

3 Definite Artikel im Japanischen

Es gibt allerdings einige Ausnahmefälle, in denen auch die japanischen Nominalphrasen definite Artikel enthalten, so daß der Parser für die betroffenen Sätze Information über Definitheit liefert:

```
[I-1, 30-31]
J: kono  jikantai wa dekireba sakete     itadakitai to omoimasu
   diese Zeit      WA möglichst freihalten FORM       TO denken
( I would like to keep this time vacant.)
```

Die definiten Artikel sind „*kono*“, „*sono*“, „*ano*“ und das Fragewort „*dono*“. Hier liefert der Parser bereits ein Analyseergebnis, das einen Eintrag für Definitheit enthält. Im von uns durchgeführten Experiment zu Terminaushandlungsdialogen treten von

[1] Die Nummerierung der Beispiele folgt der Dokumentation der Daten in Siegel et al. (1993)

insgesamt 566 Nomen 8 mit „*kono*“, 28 mit „*sono*“ und 2 mit „*dono*“, insgesamt 6,71%, auf. Die Frage des Numerus ist so jedoch noch nicht geklärt. In [I-1,30-31] spielt der Kontext eine Rolle: Wenn vorher über einen Zeitraum gesprochen wurde, ist singular angemessen, wurde aber über mehrere Zeiträume gesprochen, plural (*Diese Zeiten sollten möglichst vermieden werden*, bzw. *I would like to keep these times vacant*). Doch nicht nur definite Artikel führen zu einem Analyseergebnis mit Angaben über Definitheit oder Numerus, sondern auch weitere Artikel, Adjektive und Genitivkonstruktionen des Japanischen:

a) onaji, z.B. „*onaji shū*“ — dieselbe(n) Woche(n) (2x)

b) kazusukunai, z.B. „*kazusukunai kyūjitsu*“ — wenige Feiertage (1x)

c) tsugi, z.B. „*tsugi no hi*“ — der nächste Tag/die nächsten Tage (2x)

d) kondo, z.B. „*kondo no kaigi*“ — das nächste Treffen/die nächsten Treffen (1x)

a), c) und d) geben — wie die definiten Artikel — die Information 'definit'. b) gibt Pluralinformation an das Nomen weiter.

4 Übersetzungsäquivalente ohne Numerus– oder Definitheitsinformation

In 14,49% der Fälle in unseren Dialogdaten enthält ein (präferiertes) Übersetzungsäquivalent im Englischen kein Nomen, so daß die Suche nach Numerus- und Definitheitsinformation nicht notwendig ist. Zum einen sind das generelle Übersetzungsentsprechungen idiomatischer Äußerungen, zum anderen Realisierungen von Sprechakten in der Domäne.

Generelle Übersetzungsentsprechungen sind:

hayai jikan $\hat{=}$ early
nagai jikan/toki$\hat{=}$ long
... no hō ga tsugō ga yoi $\hat{=}$ would be better
sono hoka wa ... kanōsei nai $\hat{=}$ it would be difficult otherwise

Zum Sprechakt 'Vorschlag machen' gehören:

yotei nan desu ga (Idiom) $\hat{=}$ bei uns sieht <es> so aus (our plans are)
donna yotei ni naru ka $\hat{=}$ wie ist ...?, (how about ...)
yotei wo tatetai $\hat{=}$ schlage ich vor ... (I would propose ...)
watashidomo no yōbō $\hat{=}$ was für uns möglich ist (what is possible for us)
jikan ga toritai $\hat{=}$ wär's sehr schön bei uns (I would propose)

Zum Sprechakt 'Vorschlag beantworten' gehören:

jikan ga toreru/torenai $\hat{=}$ daran teilnehmen können (to be free)
jikan ga aru/nai $\hat{=}$ da sein (können) (be free)

Indikatoren für die Sprechakte können dazu führen, nach pragmatischen und nicht wörtlichen Übersetzungsäquivalenten zu suchen.

Bevor also versucht wird, Numerus– und Definitheitsinformation zu ergänzen, sollte geprüft werden, ob dieser Schritt überhaupt notwendig ist, da in vielen Fällen ein Äquivalent in der Zielsprache kein Nomen mehr enthält. In einzelnen Fällen wird im Übersetzungsäquivalent eine Formulierung mit zwei Nomen im Japanischen zu einem Nomen zusammengefaßt. Das betrifft zum Beispiel „*watashidomo no tokoro*“ - „*we*“. 34,63% der Nomen sind Uhrzeiten. Sie können stereotyp übersetzt werden, ohne daß Numerus oder Definitheit eine Rolle spielen, zum Beispiel: „*ichiji ni*“ – „ *at one o'clock*“.

Information über Definitheit muß nicht gesucht werden, wenn die Nominalphrase ein Genitivpronomen enthält, wie zum Beispiel „*watashitachi no tokoro no kenkyūin*“ – „*our researchers*“. Die Numerusinformation muß allerdings aus einer anderen Quelle gesucht werden. Hiervon ist nur ein geringer Teil der Nomen betroffen.

Für insgesamt 50% der Nomen aus unseren experimentell erhobenen Daten ist es daher nicht notwendig, Information über Numerus und Definitheit zu suchen, wenn zunächst nach Übersetzungsäquivalenten gesucht wird.

5 Einzigartige Entitäten, Numerale, Pluralnomen und Weltwissen

Einige Entitäten sind innerhalb der Domäne nur einmal vorhanden und bekannt. In unserer Domäne sind das die Wochentage (auch mit Vormittag oder Nachmittag: getsuyōbi no gogo - Montag Nachmittag), die Datumsangaben, die Mittagspause, das Treffen, die Firma, der Termin und das Forschungsprojekt. Diese Entitäten werden immer definit und mit Singular übersetzt. Eine domänenabhängige Transferregel für diese Entitäten muß die fehlende Information ergänzen. In unseren Daten sind 20,14kommen einzelne Auftreten von Datumsangaben, dem Treffen, der Firma, der Mittagspause, des Termins und des Forschungsprojekts. Insgesamt sind hier 26,19% der Nomen betroffen.

Eindeutige Informationen über Numerus geben die Numerale im Japanischen. Einige Beispiele aus den Daten sind:

a) ichijikan, sanjikan, nijikan

b) hitori, futari

c) hitori no hito, yonin menbā

d) kenkyūin no hitori

In diesen Fällen,7,42%, liefert der Parser Information über Numerus, so daß im lexikalischen Transfer einfache Entsprechungen möglich sind:

a) sanjikan ≙ three hours

b) hitori ≙ one colleague

c) yonin menbā ≙ four members

d) kenyūin no hitori ≙ one of the researchers

Ähnlich verhält es sich mit den Nomen „*mina*“, „*minasan*“ und „*zenin*“. In ihrer lexikalischen Information ist der Wert 'plural' bereits enthalten, wie das folgenden Beispiel zeigt:

```
[II-1, 54-56]
J: sorede . kinyōbi no gogo      da  to . ano mina jikan
   KONJ      Freitag  NO Nachmittag COP  TO       alle  Zeit
   arun de . watashidomo to shite wa . ano kinyōbi no gogo
   haben      wir                   WA       Freitag  NO Nachmittag
   ga . ichiban iin ja nai ka  to omoimasu ga
   GA    am besten    nicht   QUE TO denken
(Wenn das Freitag Nachmittag ist, haben alle Mitarbeiter Zeit, für
 uns ist Freitag Nachmittag am besten, denke (ich))
```

In 9 Fällen unserer Daten haben die Sätze die Form:

<WOCHENTAG> wa <NOMEN> desu

Zum Beispiel:

```
getsuyōbi wa kyūjitsu desu.
Montag     WA Feiertag  COP
```

Da es sich um einen Tag handelt, ist in diesem Fall singular – indefinit angemessen: „*Monday is a holiday*“. Es gibt ebenfalls zwei entsprechende Fälle mit zwei Tagen, in denen plural angemessen ist. Während dies mit Wochentagen in Subjektposition als (exakte) Regel formuliert werden kann, läßt sich allgemeiner eine Präferenzregel aufstellen, die Numeruskongruenz zwischen Subjekt und Komplement in solchen Konstruktionen bevorzugt. Daß diese Regel eine Präferenzregel und keine exakte Regel sein muß, zeigt das Beispiel:

```
[IV-2: 1-2]
J: watakushidomo no hō wa sannin     no purojekutochīmu nan desu
           wir            WA  3 Personen NO Projektteam         COP
(Wir sind ein Projektteam von drei Personen)
```

Nach den bisher beschriebenen Kriterien lassen sich für 37,1% der Nomen unserer Daten Informationen über Numerus und Definitheit finden. Für weitere 8,3% der Nomen läßt sich die Definitheitsinformation bestimmen. Für einen weiteren Fall läßt sich Numerusinformation bestimmen:

```
[I-2, 40-43]
J: kazusukunai kyūjitsu nanode
   wenige      Feiertag  weil
(Der Grund ist, daß es nur wenige Feiertage gibt)
```

50% der Nomen haben Übersetzungsäquivalente, für die keine Numerus- oder Definitheitsinformation gesucht werden muß. In 5 Fällen läßt sich nicht erkennen, welche Präferenzen für Numerus bestehen, singular wie plural ist möglich. Ein Fall läßt sich nur mit Weltwissen lösen:

```
J: kono   hi  wa shain       wa kimasen
   dieser Tag WA Mitarbeiter WA kommen(Neg)
(An diesem Tag kommen die Mitarbeiter nicht)
```

Hier ist es notwendig zu wissen, daß die Firma mehrere Mitarbeiter – und nicht nur einen – hat. Für einige wenige Nomen gibt es mit den beschriebenen Kriterien keine Lösung. Es reicht jedoch aus, hier den Default-Wert indefinit-singular anzunehmen, um keine falschen Übersetzungsäquivalente zu bekommen. Die restlichen Fälle betreffen das Nomen „*koto*". „*koto*" ist eine Nominalisierung, deren Übersetzungsmöglichkeiten in verschiedenen Kontexten an anderer Stelle genauer untersucht werden sollten. Die Äquivalente in der Zielsprache Englisch enthalten jedoch kein entsprechendes Nomen, für das Information über Numerus und Definitheit gesucht werden muß.

Für über 90% der Nomen kann so Definitheits– und Numerusinformation gefunden werden, ohne daß kompliziertere Verfahren notwendig sind.

6 Verfahren zur Bestimmung von Numerus und Definitheit

6.1 Heuristische Verfahren

Eine Möglichkeit, die Informationssuche für die übrigen Fälle zu organisieren, bilden Heuristiken. Ein Beispiel dafür ist der Ansatz von Murata/Nagao (1994). Murata/Nagao suchen nach 'Schlüsseln' in der Oberflächeninformation der japanischen Äußerungen, für die Heuristiken aufgestellt werden. Die Heuristiken für Numerus haben die Form {singular (possibility, value), plural (possibility, value), uncountable (possibility, value)}, die für Definitheit {indefinite (possibility, value), definite (possibility, value), generic (possibility, value)}. "possibility" hat den Wert 1, wenn eine Kategorie möglich ist, sonst den Wert 0. "value" ist eine relative Möglichkeit, eine Plausibilität, und nimmt Werte zwischen 1 und 10 an. Die Heuristiken werden auf eine japanische Äußerung angewandt und die "value"–Werte zusammengezählt. Das Merkmal mit dem höchsten Wert ist am wahrscheinlichsten und wird für die Übersetzung ausgewählt. Relevant für die Anwendung der Heuristiken ist Oberflächeninformation.

Zwei der von ihnen vorgestellten Heuristiken betreffen die definiten Artikel:

kono/sono/ano:

a) {indef(0,0),def(1,2),gen(0,0)}

b) {sg(1,3),pl(1,0),uncount(1,1)}

Die Heuristik a), die Definitheit bestimmt, ist nicht notwendig, wenn ein Analyseergebnis des Parsers vorliegt, das einen Eintrag 'definit' hat. Heuristik b) für Numerus ist auch für den Verbmobil–Anwendungsbereich relevant. In unseren Daten treten kono/sono/dono insgesamt 38mal auf, davon in 25 Fällen mit einem Nomen, das singular übersetzt werden muß, in 3 Fällen mit einem plural-Nomen und in 10 Fällen mit einem Nomen der Kategorie 'uncountable' (ebenfalls singular übersetzt). Die Heuristik, die Numerus betrifft, wird somit durch unsere Daten gestützt. In diesem Zusammenhang stellt sich die Frage, ob eine Heuristik derselben Art auch für „*onaji*", „*tsugi*" und „*kondo*" aufzustellen ist. In unseren Daten treten „*tsugi*" (2x), „*onaji*" (2x) und „*kondo*" (1x) nur mit Nomen im singular auf. Für die Aufstellung einer Heuristik wäre es notwendig, eine größere Menge von Daten zu untersuchen. Die anderen Heuristiken, die die Autoren vorstellen, haben für unsere Daten keine Relevanz.

6.2 Präferenzregeln

Eine andere Möglichkeit der Darstellung, wie sie auch von Schmitz/Quant (1993) vorgeschlagen wird, sind Präferenzregeln. Schmitz/Quantz stellen ein Modell vor, das hybrid in dem Sinne ist, daß soviel wie möglich exaktes Wissen mit zusätzlichem default-Wissen kombiniert wird. Dieser Ansatz ist — nach Auswertung unserer Daten — auch für das hier beschriebene System sinnvoll und effizient. Kameyama (1994) verwendet für die Formalisierung der notwendigen Präferenzregeln prioritisierte Circumscription. Damit ist es möglich, exakte Regeln und Präferenzregeln zu kombinieren. Wie dieser Formalismus für Nullpronomina genutzt werden kann, untersuchen Witt/Senf (1994). Die relevanten Regeln für Numerus und Definitheit in der VERBMOBIL–Domäne, die wir gefunden haben, lauten zusammengefaßt:

1) In der Domäne bekannte und einzigartige Entitäten werden singular - definit übersetzt. (Exakte Regel)

2) In einer Struktur <NP> wa <NP> desu besteht Numeruskongruenz zwischen der ersten und der zweiten NP. (Präferenzregel)

3) Beim Auftreten von kono,sono,dono,onaji,tsugi,kondo in einer Nominalphrase wird diese mit Singular übersetzt. (Präferenzregel)

4) Nominalphrasen werden singular-indefinit übersetzt. (Default)

Dabei hat die Regel 1 Präferenz vor Regel 2, wie das folgende (konstruierte) Beispiel zeigt:

```
kono   purojekuto to  sono purojekuto wa ohiruyasumi desu.
dieses Projekt    und jenes Projekt   WA Mittagspause COP
(Dieses Projekt und jenes Projekt haben Mittagspause)
```

„*ohiruyasumi*" sollte in diesem Beispiel mit Singular übersetzt werden, auch wenn die Nominalphrase aus einer Konjunktion besteht. Das oben zitierte Beispiel [IV-2: 1-2] zeigt ebenfalls, daß Regel 1 Präferenz vor Regel 2 hat. Ein weiteres konstruiertes Beispiel verdeutlicht die Präferenz der Regel 2 vor Regel 3:

```
kono  hi   wa kayōbi   to  mokuyōbi   desu.
diese Tage WA Dienstag und Donnerstag COP
(Diese Tage sind Dienstag und Donnerstag)
```

Obwohl „*hi*" den Artikel „*kono*" hat, muß es in einer solchen Phrase mit Plural übersetzt werden, da die zweite Nominalphrase aus einer Konjunktion besteht.

7 Schluß

Nur in wenigen Fällen, in denen auch im Japanischen definite Artikel, quantifizierende Adjektive oder Genitivkonstruktionen vorhanden sind, kann ein Ergebnis des Parsers Information über Definitheit oder Numerus enthalten. Im 50% der Fälle unserer Daten kann jedoch eine Suche nach der relevanten Information vermieden werden, wenn zunächst geprüft wird, ob ein Äquivalent in der Zielsprache vielleicht kein Nomen enthält. Für die anderen Fälle wurden Kriterien zur Suche nach Numerus– und Definitheitsinformation vorgestellt. Diese Verfahren sind stark an den Erfordernissen der Domäne ausgerichtet und ihre Bedeutung anhand natürlicher Daten quantifiziert. Bei der Anwendung auf Dialoge spielen generelle Verfahren eine untergeordnete Rolle gegenüber domänenabhängigen. Die Beobachtungen natürlicher Daten bestätigen den 'Sublanguage'-Ansatz: "If a source language analyzer is based on a sublanguage grammar, instead of (or in addition to) a grammar of the 'whole' language, then a significant gain in efficiency is possible." (Kittredge 1987:63).

Die gefundenen Regeln können als Heuristiken oder Präferenzen dargestellt werden. Dabei sollten - wie auch bei Schmitz/Quantz (1993) vorgeschlagen ist, Präferenzregeln mit exakten Regeln kombiniert werden. Welche Formalisierung für VERBMOBIL vorzuziehen ist, ergibt sich aus der weiteren Entwicklung der Formalismen.

Literatur

Kameyama, Megumi (1994). *Indefeasible Semantics and Defeasible Pragmatics.* Technical Note 544. SRI International, Menlo Park, CA.

Kittredge, Richard I. (1987): The Significance of Sublanguage for Automatic Translation. In: Nirenburg, Sergei (ed.): *Machine Translation.* Cambridge: Cambridge Univ. Press: 136-144.

Murata, Masaki & Nagao, Makoto (1993): Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In: *Proceedings of the fifth International Conference on Theoretical and Methodological Issues in Machine Translation*: 218-225.

Schmitz, Birte & Quantz, J. Joachim (1993): *Defaults in Machine Translation.* KIT Report 106. Technische Universität Berlin.

Siegel, Melanie (1993). *Dialogdolmetschen. Eine Pilotstudie zu aufgabenorientierten Dialogen (Terminabsprachen) Japanisch-Deutsch. Arbeitsberichte Computerlinguistik 3-93.* Universität Bielefeld.

Siegel, M., H. Kuroda & E. Kubo. 1993. *Dialogdaten: Terminplanung Japanisch-Deutsch.* Arbeitsberichte Computerlinguistik 2-93, Universität Bielefeld.

Witt, Andreas & Senf, Thomas (1994): *Formalisierung von Kontext und sprachlichem Wissen mit Prioritisierter Circumscription.* Verbmobil-Memo 55. Bielefeld: Universität Bielefeld.