ZUR KLASSIFIKATION VON FRAGEN UND NICHT-FRAGEN ANHAND INTONATORISCHER MERKMALE

A. Batliner[1], E. Nöth[2], R. Lang[2], G. Stallwitz[2]

[1] Institut für Deutsche Philologie, Universität München

[2] Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen

1. Fragestellung

Bei der Kennzeichnung von Fragen vs. Nicht-Fragen können intonatorische sowie nicht-intonatorische Merkmale (syntaktische, semantische, etc.) eine Rolle spielen. Dieser Beitrag untersucht an einem großen Korpus, inwieweit der Grundfrequenzwert am Äußerungsende als Frage-/Nicht-Frageindikator ausreicht, wie dieser Wert aussehen soll (Rohwert in Hz oder umgerechnet zu Bezugsgrößen), und ob es sprecherspezifische Strategien beim Einsatz dieses Parameters gibt.

2. Korpora

Die unserer Stichprobe zugrundeliegenden Korpora bestehen aus Minimalpaaren und -tripeln, bei denen grundsätzlich die Intonation die Satzmodus- bzw. Satzfokus-Unterscheidung gewährleistet (Satztyp bzw. Satzakzent). Ein Überblick über das Satzmodus- und Satzfokus-Modell findet sich in /2,5/. Zwei Sprechergruppen, bestehend aus je drei weiblichen und drei männlichen Sprechern, produzierten 155 in einen modus- und fokussteuernden Kontext eingebettete Testsätze. Jeder Testsatz wurde zwei- bis viermal gesprochen (2074 Äußerungen). Die intonatorischen Kennwerte wurden an Mingogrammen bestimmt und mit dem Statistikpaket SPSSPC+ ausgewertet. Da in 75 Fällen die Sprecher laryngalisierten und deshalb der Grundfrequenzwert nicht ermittelt werden konnte, blieben 1999 Äußerungen in der Analyse. Die vier Teilkorpora K1 bis K4 werden in Tab. 1 kurz vorgestellt. Die Korpora K3 und K4 lagen auch in digitalisierter Form vor, so daß für diese Korpora sämtliche Experimente zusätzlich mit automatisch extrahierten Kennwerten durchgeführt wurden. Für diese Experimente wurden die Laryngalisierungen (2% der Fälle) nicht ausgesondert. Dies entspricht der Vorgehensweise in einem automatischen Erkennungssystem.

Korpus	Sprecher-gruppe	Test-sätze	An-zahl	syntakt. Struktur	Minimal-paare	beschrie-ben in
K1	1	71	896	Verb-Erst- Verb-Zweit-	Modus	/3,5/
K2	2	45	573	Verb-Letzt-	Modus	/2/
K3	2	26	355	Verb-Erst- Verb-Zweit-	Modus+ Fokus	/2/
K4	2	13	175	Verb-Erst- Verb-Zweit-	Modus+ Fokus	/2,4/

Tab. 1: Teilkorpora der Gesamtstichprobe

3. Lernstichproben und Merkmalsatz

Mit der Diskriminanzanalyse wurden die folgenden Konstellationen berechnet: Als **Prädiktorvariablen** der F_0-Wert am Äußerungsende (*Off*), der F_0-Wert am Äußerungsanfang (*Ons*), der größte F_0-Wert (*Max*) und der kleinste F_0-Wert (*Min*). Als **Transformationen** F_0-Rohwerte (*Hz*), F_0-Werte umgerechnet in Halbtonwerte (*Ht*); *Hz*- und *Ht*-Werte minus sprecherspezifischem Basiswert (*Hzbas* bzw. *Htbas*); *Hz*- und *Ht*-Werte minus einen Äußerungsmittelwert aus den vier Werten *Ons*, *Off*, *Max* und *Min* (*Hzmit* bzw. *Htmit*). Als **Lernstichproben** alle Sprecher (*ln*); n-1 Sprecher (*ln-1*), also leave-one-out zur Simulation der Sprecherunabhängigkeit; ein Sprecher (*l1*) zur Generalisierung von einem auf andere Sprecher. Die im weiteren besprochenen Analysen werden wie folgt gekennzeichnet: '*ln/Off/Htbas*' steht für eine Diskriminanzanalyse, bei der Lern- gleich Prüfstichprobe ist, und als Prädiktorvariable der Offset in Halbtönen umgerechnet zum sprecherspezifischen Basiswert benutzt wurde.

Transformationen: Die im allgemeinen als **'gehörsadäquat'** angesehene Transformation der *Hz* in *Ht*-Werte zur Normierung des unterschiedlichen **Stimmumfangs** (Range) von Männern und Frauen ergab eher schlechtere Ergebnisse. Bei *Hzmit* und *Htmit* sind die Unterschiede zu vernachlässigen (siehe Tab. 2). Der Range betrug bei den Frauen im Mittel 146 *Hz* bzw. 10.3 *Ht*, bei den Männern 91 *Hz* bzw. 11.7 *Ht*. In *Hz* ist der Range der Frauen im Mittel also größer als der der Männer, in *Ht* ist es umgekehrt. Möglicherweise liegt eine adäquatere Transformation **zwischen** den *Hz*- und *Ht*-Werten.

Eine Normierung der **Stimmlage** zu einem Bezugswert, egal ob zum Basis- oder Mittelwert, ergab dagegen immer bessere Ergebnisse als bei den *Hz*- oder *Ht*-Werten (siehe Tab. 2). Der Grund wird klar, wenn man z.B. *Ht* und *Htbas* bei *ln-1/Off* miteinander vergleicht und das Ergebnis nach Fragen (F) und Nicht-Fragen (N) sowie nach männlichen (M) und weiblichen Sprechern (W) aufschlüsselt (Tab. 3): Wegen der unterschiedlichen Stimmlagen werden ohne die Normierung bei den Frauen viele Nicht-Fragen fälschlich als Fragen klassifiziert und umgekehrt bei den Männern viele Fragen als Nicht-Fragen.

Prädiktorvariablen: Bei *Hz* und *Ht* verbesserte zwar die Hinzunahme weiterer Prädiktoren die Klassifikation, da damit automatisch die Stimmlage mit in die Berechnung einging, bei

Prädiktor-Variable	Lernstichprobe ln	ln-1	l1
Off/Hz	75.4	70.7	65.2
Off/Ht	67.9	67.0	64.9
Off/Hzmit	86.9	86.9	86.3
Off/Htmit	86.8	87.1	86.2
Off/Htbas	86.6	86.7	85.7

Tab. 2: Erkennungsraten für unterschiedliche Lernstichproben

		F Ht	F Htbas	N Ht	N Htbas
F	W	92.4	77.5	7.6	22.5
	M	40.7	82.9	59.3	17.1
N	W	45.5	9.8	54.5	90.2
	M	1.3	8.2	98.7	91.8

Tab. 3: Erkennungsraten für Fragen und Nicht-Fragen bei *ln-1/Off/Ht* und *ln-1/Off/Htbas*

Htbas und *Htmit* war die Verbesserung aber minimal und immer unter 1%. Zum einen liegt das an der Korrelation der Variablen untereinander, zum anderen daran, daß diese anderen Variablen für die Unterscheidung von Frage/Nicht-Frage kaum relevant sind. (Das gilt nicht für die Unterscheidung anderer Satzmodi, vgl. /3/.)

Lernstichproben: Tab. 2 zeigt, daß der Unterschied bei *Off* zwischen *ln*, *ln-1* und *l1* bei *Hz* jeweils 5% beträgt, bei *Hzmit*, *Htmit* und *Htbas* aber vernachlässigbar ist; die einzelnen Sprecher wenden also grundsätzlich die gleiche Strategie an (vgl. aber weiter unten).

Automatische vs. handextrahierte Parameter: Die Erkennungsraten für die automatischen Paramter (z.B. *ln-1/off/htmit:* 90.7% für K3 und 88.3% für K4) sind zwar etwas niedriger, aber mit denen für handextrahierte Parameter (92.8% für K3 und 93.2% für K4) durchaus vergleichbar.

4. Fehleranalyse

Bild 1 zeigt die Verteilung in % der *Off/Htbas*-Werte für Fragen und Nicht-Fragen. Für die Fragen zeigt sich eine klare bimodale Verteilung, wobei der kleinere Gipfel im Zentrum der Nicht-Fragen liegt. Im Gegensatz dazu ist die Verteilung der Nicht-Fragen unimodal, aber rechtsschief und reicht nicht ins Zentrum der Fragen.

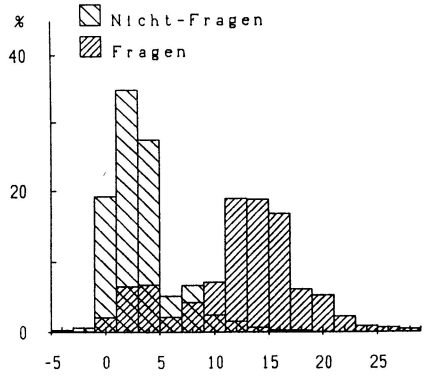

Bild 1: *Off/Htbas*-Werte für Fragen und Nicht-Fragen

Durchschnittlich 12 Hörer klassifizierten alle Äußerungen ohne Kontext nach Satzmodus (vgl. /5/). Diese Klassifizierung wurde umkodiert in einen Wert zwischen 0.0 und 1.0 für Frage/Nicht-Frage. Wie bei der Wahrscheinlichkeit der Gruppenzugehörigkeit, die sich aus der Diskriminanzanalyse ergibt, wurden nun die Werte unter 0.5 als 'falsch' und die über 0.5 als 'richtig' bezeichnet. Damit ergibt sich die Aufteilung der 1999 Fälle in Tab. 4.

Tab. 4: Klassifikation durch Hörer und Diskriminanzanalyse

Hörer-urteil	Diskriminanzanalyse (ln/Off/Htbas) richtig		falsch	
richtig	I	1625 (81.3%)	II	221 (11.1%)
falsch	III	105 (5.3%)	IV	47 (2.4%)

Unser einfaches Modell konvergiert in gut 80% der Fälle mit der Bewertungsstrategie der Hörer (Gruppe I). Eine genauere Diskussion von Gruppe III und IV muß unterbleiben (vgl. dazu aber /5,2/ und unten); wir konzentrieren uns in der folgenden Diskussion auf die häufigsten Fälle aus Gruppe II, da hier anzunehmen ist, daß andere Merkmale als der

Offset die Klassifikation bestimmen. Die meisten dieser Fälle sind für den linken Gipfel bei der bimodalen Verteilung der Fragen verantwortlich, weisen also einen tiefen Offset auf.

47 Fälle (2.4%) sind Alternativfragen wie *Möchten Sie Mohn oder Streusel?*, die bei expliziter Angabe aller Wahlmöglichkeiten regulär mit tiefem Offset produziert werden. In diesen wie in ähnlichen Fällen disambiguieren also **nicht-intonatorische Merkmale** wie Verbstellung und Verbsemantik.

Ein Blick auf die einzelnen Sprecher bei *ln-1/Off/Htbas* zeigt, daß bei der Prädiktion der Fragen eine Sprecherin deutlich schlechter abschneidet (48%) als alle anderen Sprecher, die zwischen 75% und 93% liegen. Wenn der Fokus nicht auf der letzten Phrase liegt, sondern auf der vorletzten, werden Frage und Fokus normalerweise durch einen F_0-Abfall und einen ausgeprägten F_0-Anstieg auf dieser Phrase indiziert; ein hoher finaler Offset ist dann fakultativ und 'normal', wird aber von dieser einen Sprecherin nicht realisiert (20 Fälle aus K3). Eine Modusindizierung durch solche **anderen intonatorischen Merkmale** mag zwar sprecherspezifisch sein, ist aber dennoch regulär.

In einigen Fällen, besonders aus Gruppe IV, muß der **Kontext** allein disambiguieren, so z.B. bei *Schlafen Sie ?/!* (Frage oder Imperativ). Hier ist auch bei Fragen ein tiefer Offset regulär.

5. Relevanz für die automatische Spracherkennung (ASE)

Fragen sind dialogsteuernd und deshalb von unmittelbarem Interesse für die ASE. Die untersuchten Korpora sind nach linguistischen Gesichtspunkten konstruiert und geben sicherlich die Verhältnisse einer Mensch-Maschine-Interaktion (MMI) nicht wieder. Auf der anderen Seite besitzen sie den Vorteil der Vollständigkeit: Anhand des Materials läßt sich z.B. zeigen, daß es Frage-Klassen gibt, die in der MMI durchaus vorkommen können (z.B. Alternativfragen und W-Fragen), die systematisch keine prototypische Fragekontur besitzen und mit nicht-intonatorischen Mitteln klassifiziert werden müssen. Der zu Bezugswerten umgerechnete Offset hat sich als ein stabiles Merkmal erwiesen - allerdings nicht als das einzig relevante: Für eine optimale Erkennung müssen die anderen erwähnten intonatorischen und nicht-intonatorischen Merkmale mit berücksichtigt werden. Unter Umständen ist der Kontext, in dem die Äußerung steht, unbedingt notwendig.

Die Arbeiten an der Universität München entstanden im Rahmen des DFG-Projekts 'Modus-Fokus-Intonation', die an der Universität Erlangen im Rahmen des BMFT Verbundvorhabens 'Sprachverstehende Systeme'.

/1/ Altmann, H. (Hg.): "Intonationsforschungen", Max Niemeyer Verlag, Tübingen, 1988.
/2/ Altmann, H., Batliner, A., Oppenrieder, W. (Hgg.): "Modus - Fokus - Intonation", Max Niemeyer Verlag, Tübingen, erscheint voraussichtlich 1989.
/3/ Batliner, A.: "Produktion und Prädiktion. Die Rolle intonatorischer und anderer Merkmale bei der Bestimmung des Satzmodus", in /1/, S.207-221, 1988.
/4/ Nöth, E., Batliner, A., Lang, R., Oppenrieder, W.: "Automatische Grundfrequenzanalyse und Satzmodusdifferenzierung", in Tillmann, H.G., Willée G. (Hgg.): "Analyse und Synthese gesprochener Sprache", S.59-66, Georg Ohms Verlag, Hildesheim, 1987.
/5/ Oppenrieder, W.: "Intonatorische Kennzeichnung von Satzmodi", in /1/, S.169-205, 1988.