Institut für Deutsche Philologie
Ludwig-Maximilians-
Universität

Prof. Dr. H. Altmann

Schellingstr. 3
D–8000 München 40
(089) 2180–2916

Verbundprojekt ASL
- Südverbund -

Architekturen von Systemen
zur integrierten Analyse von
Sprachlauten und
Sprachstrukturen

Lehrstuhl für Informatik 5
(Mustererkennung)
Universität Erlangen-Nürnberg

Prof. Dr.-Ing. H. Niemann
Dr.-Ing. E. Nöth

Martensstr. 3
D–8520 Erlangen
(09131) 85–7774

Die prosodische Markierung des Satzmodus in der Spontansprache

Methodologie und erste Ergebnisse

Anton Batliner, Andreas Kießling, Elmar Nöth

ASL-Süd—TR—14—93/LMU

Februar 1993

Gehört zum Antragsabschnitt: 4.6 Prosodie

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter den Förderkennzeichen 01IV102H0 und 01IV102F4 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

"You crazy," said Max.
It was either a statement
or a question.
(John le Carré: Tinker tailor soldier spy)

1 Einleitung

Die bisherige Forschung zur prosodischen Markierung des Satzmodus arbeitete fast ausschließlich mit konstruiertem, elizitiertem Material. Man war daher letztlich auf die Vermutung angewiesen, daß die Ergebnisse auf die Spontansprache übertragbar sind. Darüber hinaus ist uns – zumindest für das Deutsche – keine Arbeit bekannt, in der die prosodische Markierung der in spontaner Sprache häufig vorkommenden unvollständigen (elliptischen) Strukturen Thema ist. Die vorliegende Arbeit konzentriert sich daher auf diese beiden Punkte: auf den Vergleich von spontaner und (parallelisierter[1]) gelesener Sprache, sowie auf den Vergleich vollständiger, satzförmiger Strukturen mit elliptischen, aber gleichermaßen satzwertigen Strukturen. Das Hauptgewicht liegt auf der Frage/Nicht-Frage-Unterscheidung. Die dabei auftretenden methodologischen Probleme und die unterschiedlichen Ansätze zu ihrer Bewältigung werden besprochen; zum allgemeinen Unterschied in der prosodischen Kennzeichnung von spontaner und gelesener Sprache vgl. [BJKN92].

2 Material und experimentelles Design

Sprecher waren vier Studenten (3 weibliche: C, X und A, ein männlicher: F), die jeweils paarweise an einer Sitzung teilnahmen; die Sprecher innerhalb eines Paares (C und X bzw. A und F) waren miteinander befreundet. Die zwei Sprecher saßen sich ohne Blickkontakt in einem Versuchsraum des Psychologischen Instituts in München gegenüber und gaben sich gegenseitig Anweisungen, was der Partner mit auf dem Tisch aufgebauten Klötzchen machen sollte. Die Sitzungen enthielten jeweils drei Durchgänge mit unterschiedlichen Aufgaben und dauerten, inkl. Pausen, ca. zwei Stunden. Die Partner hatten immer die gleiche Anzahl Holzklötzchen oder Baukörper. Im ersten Durchgang mußte der eine Partner das vor ihm aufgebaute Dorf in die vom anderen Partner beschriebene Stadt umbauen, im zweiten Durchgang wurden die Aufgaben vertauscht. Im dritten Durchgang wurden Klötzchen als unterschiedlich gefährliche Tiere definiert und abwechselnd in eines von fünf Gattern gelegt; die Partner mußten sich darüber einigen, ob eventuell schon vorhandene Tiere von den Neuzugängen gefressen werden oder nicht. Die Aufgaben waren so angelegt, daß sich kurze Klärungsdialoge mit häufigem Sprecherwechsel ergaben, nicht längere, raisonierende Passagen o.ä. Im Gegensatz etwa zu einem erzählenden Monolog oder zu einem freien Vortrag mit längeren Planungspausen ergab sich eine "echt" spontane, lebhafte Unterhaltung, ohne daß den Versuchspersonen bewußt war, daß ihre Sprache und nicht, wie ihnen gesagt wurde, ihr kooperatives Verhalten untersucht wurde.

[1] Die spontan gesprochenen Äußerungen wurden transliteriert und den einzelnen Sprechern bzw. ihren Partnern noch einmal zum Lesen vorgelegt.

	Gesamt	ausgewählt
nicht elliptische Fragen	37	24
elliptische Fragen	46	28
nicht elliptische Aussagen	179	15
elliptische Aussagen	135	17
nicht elliptische Imperative	6	6
elliptische Imperative	2	2
nicht elliptische Exklamative	0	0
elliptische Exklamative	0	0

Tabelle 1: Gegenüberstellung von vorhandenen und ausgewählten Äußerungen eines Durchgangs der Sprecherinnen C und X, aufgeteilt nach Satztypklasse (Fragen, Aussagen, Imperative, Exklamative) und Vollständigkeit (Nicht-Ellipsen vs. Ellipsen).

Domäne der Untersuchung war die Satzebene, die Textebene blieb ausgeklammert. Aus Aufwandsgründen konnte nicht das gesamte Material (ca. 5 Stunden Sprachmaterial inklusive Pausen) bearbeitet werden. Auswahlkriterium war zum einen, daß wir nur satzwertige Äußerungen ohne Häsitationen, Abbrüche, Neuansätze etc. aufnahmen; zum anderen wurden zu leise oder von anderen Geräuschen bzw. vom Partner überlagerte Äußerungen ausgesondert. Der entscheidende Grund für den Ausschluß von Häsitationssphänomenen u.ä. war, daß ein paralleles Lesekorpus bei "naiven", d.h. nicht geschulten Sprechern nur sinnvoll ohne Häsitationen etc. erstellbar ist, da die gezielte Produktion solcher spontansprachlicher Phänomene diesen Sprechern Schwierigkeiten bereiten dürfte.

Die Äußerungen wurden klassifiziert nach Satztypen (Feinkategorien nach dem Satzmodussystem von [Alt87] sowie Fragesätze und Nicht-Fragesätze als Grobkategorien) und nach syntaktischer Vollständigkeit (Nicht-Ellipsen vs. Ellipsen). Wie zu erwarten, bestand der überwiegende Teil der Äußerungen aus Aussagen; bei Aussagen wurden die Auswahlkriterien (Signalqualität, Sprecherüberlagerung, etc.) daher strenger gehandhabt, um die Zahl der zu bearbeitenden Äußerungen im Rahmen zu halten; bei den anderen, prosodisch "interessanteren" Satztypen versuchten wir dagegen, möglichst alle Äußerungen aufzunehmen, die den angeführten Kriterien entsprachen.

Nach ca. 9 Monaten lasen die Sprecher in einer Einzelsitzung die so ausgewählten Äußerungen, und zwar sowohl die eigenen als auch die des jeweiligen Partners. Die Äußerungen waren in einen genügend großen Kontext eingebettet und wurden in schriftlicher Form vorgelegt. Sie wurden allerdings nicht in der orthographisch korrekten, "kanonischen" Form vorgegeben, sondern – ohne Satzzeichen – in gemäßigter Umgangssprache, angepaßt an das spontansprachliche

Pendant, wie etwa *"also was ham ma jetz für Steine"* statt *"also was haben wir jetzt für Steine"* oder gar *"also welche Steine haben wir jetzt"*. Damit war gewährleistet, daß die Lesevorlage der Spontanäußerung sehr nahekommt (möglichst gleiche Anzahl von Segmenten und damit gleiche Silbenanzahl) und somit direkt mit ihr verglichen werden kann. In diesem Leseregister fehlt also das Umsetzen der kanonischen Schriftsprache, es fehlen aber nicht die anderen lesetypischen Planungs- und Umsetzungsprozesse. (Es ist noch nicht klar, ob dieser Unterschied einen Vergleich mit bisherigen, an gelesenem Material gewonnenen Ergebnissen tangiert.)

Um den Aufwand im Rahmen zu halten, wurde nur für einen der drei Durchgänge des Sprecherpaares C und X eine Statistik über die linguistische Klassifikation aller Äußerungen, auch der mit eindeutig schlechter Qualität, aufgestellt. (Bei den anderen Durchgängen wurden nur die Äußerungen mit guter Qualität linguistisch klassifiziert.) Tab. 1 gibt für diesen Durchgang eine Gegenüberstellung zum Verhältnis von vorhandenen und ausgewählten Satzmodi, aufgeteilt nach Satztypklasse (Fragen, Aussagen, Imperative, Exklamative) und Vollständigkeit (Nicht-Ellipsen vs. Ellipsen). Besonders bei den nicht ausgewählten, häufig unvollständigen Äußerungen ist natürlich die Zuordnung wegen des Fehlens satzmodusindizierender Merkmale nicht immer eindeutig möglich; diese Unsicherheit dürfte aber am globalen Verhältnis der Satztypen untereinander nicht viel ändern.

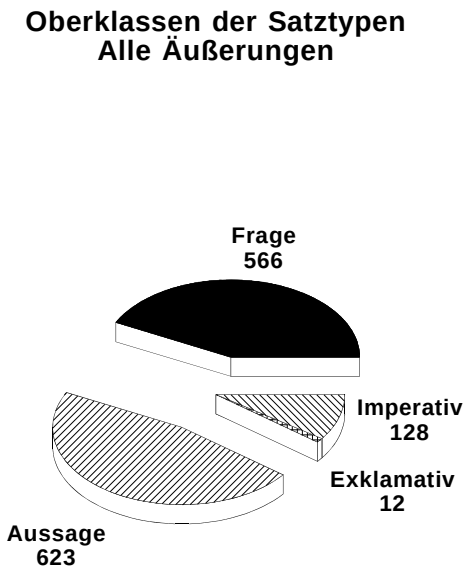

Abbildung 1

Das Kuchendiagramm in Abb. 1 zeigt die Verteilung der Satztypklassen für das gesamte ausgewählte Korpus (spontane und gelesene Äußerungen), Abb. 2 zeigt die Verteilung von Nicht-Ellipsen und Ellipsen pro Satztypklasse; die Zuweisung zur jeweiligen Satztypklasse basiert auf der weiter unten besprochenen linguistischen Klassifikation. Ein Vergleich mit Tab. 1

zeigt, daß die Auswahl im großen und ganzen repräsentativ ist – wie gesagt, mit der einen Ausnahme, daß Aussagen unterrepräsentiert sind.

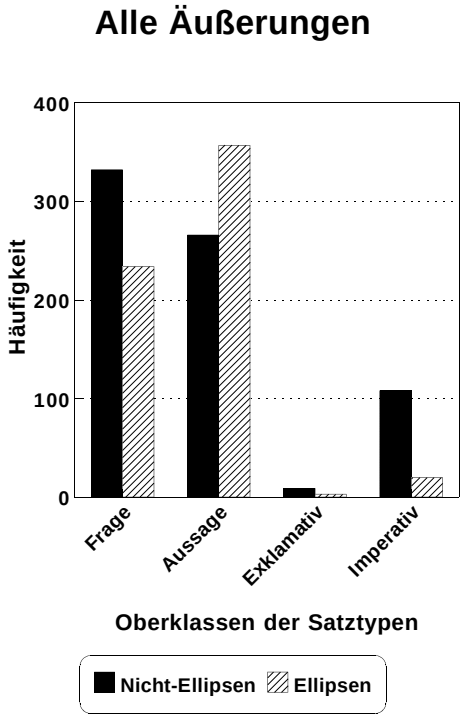

Abbildung 2

Die Aufnahmebedingungen entsprachen in beiden Sitzungen denen einer ruhigen Büroumgebung. Die Äußerungen wurden mit 12 Bit Auflösung und einer Abtastfrequenz von 10 kHz digitalisiert. Bei den spontanen Äußerungen und den gelesenen Äquivalenten zu den eigenen Äußerungen sowie zu denen des Partners handelt es sich insgesamt um 1329 Fälle (ca. 28 Minuten Sprachmaterial, Zahl der Äußerungen: C:480, X:417, A:210, F:222). Von der Verarbeitung seien hier nur die Schritte erwähnt, mit denen die weiter unten beschriebenen Merkmale extrahiert wurden: Eine berechnete Stimmhaft/stimmlos-Entscheidung wurde manuell anhand des Zeitsignals, gegebenenfalls auditiv unterstützt, korrigiert. Mit Hilfe dreier automatischer $F0$-Verfahren (vgl. [BKKN91, Hes91, Ree89, KKN*92]) wurden $F0$-Konturen ermittelt, aus denen eine Referenzkontur (je Frame, d.h. alle 12.8 msec., ein Hertzwert) durch auditiv unterstützte Handkorrektur erstellt wurde. In kritischen Bereichen wurde der $F0$-Wert periodengenau am Zeitsignal ermittelt. Irreguläre Passagen (laryngalisierte Bereiche) wurden gehörsadäquat interpoliert (vgl. [BKKN91]); $F0$-Sprünge, die sich als Artefakte der Verfahren herausstellten, wurden geglättet.

3 Zur Problematik der Satzmodusbestimmung

Das Arbeiten mit konstruiertem und elizitiertem Material hat einen nicht zu unterschätzenden heuristischen Vorteil gegenüber dem Arbeiten mit spontanem Material: Der Satzmodus der zu untersuchenden Äußerungen ist "via Setzung" vorgegeben, z.B. durch eine sorgfältige Konstruktion des Kontextes, durch eine explizite Instruktion der Sprecher oder einfach durch die Zeichensetzung. Nach der Aufnahme ist es "nur" noch nötig, Fehlproduktionen der Sprecher auszusondern; dies kann z.B. mittels eines Hörtests mit "naiven" Versuchspersonen geleistet werden. Eine annähernd eindeutige Klassifizierung ist dagegen bei spontaner Sprache nicht möglich. Dies gilt insbesondere für die in spontaner Sprache häufig auftretenden elliptischen Strukturen, da hier oft die sonst vorhandenen syntaktischen Satzmodusindikatoren nicht existieren; unsere Hypothese ist, daß in diesen Fällen der intonatorischen Markierung eine größere Rolle zukommt. Es ist aber methodisch nicht sehr befriedigend, sondern schon fast eine Art von "selbsterfüllender Prophezeihung", wenn das zu untersuchende Merkmal in die anfängliche Klassifizierung als entscheidendes Kriterium mit eingeht. Hinzu kommt, daß eine einfache dichotome Klassifizierung (+/- Frage) zu kurz greifen dürfte; wir nehmen an, daß es in der Intention des Sprechers Abstufungen der Fragehaltigkeit gibt, und daß, dadurch bedingt, die Antwortobligation des Hörers mehr oder weniger stark ausgeprägt sein kann. Ein Patentrezept für dieses Klassifikationsdilemma gibt es nicht[2]. Wir haben uns daher dafür entschieden, drei verschiedene Klassifikationsdurchgänge durchzuführen:

1. **Linguistische (grammatische) Klassifikation:** Ein linguistischer Experte klassifizierte die Äußerungen anhand des Satzmodussystems von [Alt87].

2. **Kontextklassifikation:** Die Äußerungen wurden anhand kontextueller Kriterien klassifiziert; vgl. unten.

3. **Hörerklassifikation:** In einem Hörtest klassifizierten jeweils 10 "naive" Versuchspersonen die Äußerungen nach ihrer Satztypklasse als Frage, Aussage, Exklamativ oder Imperativ.

Am Anfang stand die linguistische Klassifikation des spontansprachlichen Materials, die – neben der signalphonetischen Qualität – Grundlage für die Auswahl des Lesematerials bildete. Die gelesenen Äußerungen wurden später ebenfalls linguistisch klassifiziert. Es folgten, zeitlich überlappend und voneinander unabhängig, die Kontext- und die Hörerklassifikation.

Die linguistische Klassifikation folgt formalen, nicht funktionalen Kriterien. Die beiden anderen Klassifikationsarten folgen funktionalen, keinen formalen Kriterien. Man beachte, daß diese drei Klassifizierungsmöglichkeiten keine Alternativen sind, sondern sich ergänzen. Die linguistische Klassifikation mit ihren formalen Kriterien ist die einzige, die sich direkt in ein linguistisches Beschreibungssystem einbetten läßt; sie berücksichtigt sowohl prosodische Merkmale als auch die anderen grammatischen Merkmale (kategoriale Füllung, Verbstellung und Imperativmorphologie). Wenn man – wie bei der Kontextklassifikation – einen Dialog nur geschrieben vor sich hat, ist die Klassifizierung mithilfe prosodischer Merkmale natürlich nicht

[2]Es wäre problematisch, bei einem spontansprachlichen Korpus "Fehlproduktionen" mit einem Hörtest auszuschalten, wie man das bei gelesener Sprache machen kann. Sie sind natürlich genauso möglich, aber die Grenze zwischen ambigen, mißverständlichen oder vom Partner einfach nicht verstandenen Äußerungen auf der einen Seite und "echten" Fehlproduktionen auf der anderen Seite ist nicht eindeutig festzulegen.

möglich – und in unserem Fall auch nicht wünschenswert. Die Hörerklassifikation ("*out of the blue*"–Sätze) entspricht der Aufgabe eines Prosodiemoduls in der automatischen Spracherkennung mit Rekursmöglichkeit auf andere, höhere Ebenen, aber ohne Dialogwissen; komplementär dazu baut die Kontextklassifikation auf dem Dialogwissen auf, verwendet aber keine prosodische Information. Eine inhärente Schwierigkeit der linguistischen und der Kontextklassifikation ist es, daß aus Aufwandsgründen die Arbeit jeweils nur von einem einzigen Experten geleistet werden konnte; damit sind ein individueller Bias und/oder Flüchtigkeitsfehler nicht auszuschließen. Darüber hinaus beruht die Kontextklassifikation insbesondere bei den nicht eindeutigen Fällen notwendigerweise auf einem "Außenkriterium": auf der "sichtbaren" Reaktion des Partners, die auf eine Interpretation des Partners bzgl. der Intention des Sprechers zurückzuführen ist. Der Partner kann aber auch z.B. auf eine von ihm klar als Aussage interpretierte Äußerung verneinend reagieren – und sie damit im Sinn der Kontextklassifikation als mögliche Frage kennzeichnen. Ein Beispiel zeigt der folgende Dialog:

Sprecher 1: *"Der grüne Klotz ist auf dem roten."*
Sprecher 2: *"Nein, das stimmt nicht."*

Die Wortstellung des ersten Satzes ist die einer Aussage; ob er mit einer "Frageintonation" (steigender $F0$-Verlauf) geäußert wurde, kann bei der Kontextklassifikation nicht festgestellt werden. Die verneinende Reaktion von Sprecher 2 deutet darauf hin, daß die Äußerung von Sprecher 1 eine Frage sein könnte; man kann aber praktisch auf jede Aussage mit einer Verneinung reagieren. Der Satzmodus des ersten Satzes kann deshalb nicht eindeutig festgelegt werden: Es handelt sich also um eine "mögliche Frage".

Ziel der Kontextklassifikation ist es, den Funktionstyp der spontansprachlichen Äußerungen zu bestimmen, ohne intonatorische Merkmale heranzuziehen. Diese Klassifikation bestimmt die Fragehaltigkeit der Äußerungen nach inhaltlichen Kriterien und nach der Dialogstruktur (welche Äußerungen gehen voraus, welche Reaktion folgt?). Dabei werden die Äußerungen zum einen in ihrem dialogischen Zusammenhang untersucht (Geht eine Frage des Partners voraus? Folgt eine Antwort des Partners?) und zum anderen in ihrem monologischen Zusammenhang (Geht eine selbst gestellte Frage voraus, die sich der Sprecher selber zu beantworten sucht? Folgt eine eigene Antwort? Wiederholt der Sprecher die vom Partner gegebene Antwort?)

Alle Äußerungen, auf die keine bestätigende, verneinende, präzisierende, oder inhaltlich beantwortende Reaktion folgt, werden nicht als potentielle Fragen und damit als Nicht-Fragen eingestuft. Stammt eine Äußerung von einem Sprecher, der sich (hinsichtlich des Inhalts dieser Äußerung) im Gegensatz zum Partner in der Rolle des Wissenden befindet und dies auch weiß, so ist diese Äußerung ebenfalls eine Nicht-Frage. Äußerungen, auf die eine Antwort erfolgt, sind keine Nicht-Fragen. Wir verwenden im folgenden eine Kombination aus dialogischer und monologischer Klassifikation mit den vier Klassen:

Nicht-Frage, mögliche Frage, wahrscheinliche Frage, Frage.

Die beiden Typen Nicht-Frage und Frage beruhen auf eindeutigen Zuweisungen, die beiden mittleren nicht. Die Kontextklassifikation wurde nur für die spontanen Äußerungen durchgeführt und automatisch auf die gelesenen Pendants übertragen, bei denen ja der gleiche funktionsindizierende Kontext vorgegeben war. Im folgenden steht SPONTAN abkürzend für die

nicht-gelesenen "spontanen Äußerungen" bzw. für "Spontanregister", und ebenso GELESEN für die "gelesenen Äußerungen" bzw. für "Leseregister".

Da die Hörerklassifikation von keinem möglichen Expertenbias belastet und auch intersubjektiv abgesichert sind, werden wir im folgenden grundsätzlich die auf ihnen beruhende Klassifizierung als Referenz betrachten. Hinzu kommt, daß allein eine solche Hörerklassifikation durch "naive" Versuchspersonen theorieunabhängig ist und demnach sinnvollerweise als gemeinsame Beschreibungsplattform angesehen werden sollte.

4 Ergebnisse

Die im weiteren Verlauf diskutierten Ergebnisse und Zahlen beziehen sich im Normalfall auf alle 1329 Äußerungen. Wenn eine Teilmenge betrachtet wird, so ist dies jeweils gesondert angegeben. Abb. 3 zeigt für SPONTAN das Auftreten der vier Kontextklassen, aufgeteilt nach Nicht-Ellipsen und Ellipsen. Ellipsen unterscheiden sich systematisch von Nicht-Ellipsen: da bei ihnen öfters satzmodusindizierende Merkmale fehlen, ist bei der Kontextklassifikation, die ja die prosodischen Merkmale nicht mitberücksichtigt, eine Zuweisung zu den eindeutigen Klassen Nicht-Frage bzw. Frage schwieriger; es gibt also weniger eindeutige elliptische Nicht-Fragen und insbesondere Fragen als nicht-elliptische. Bei den nicht eindeutigen Klassen "mögliche Frage" bzw. "wahrscheinliche Frage" kehrt sich das Verhältnis von Ellipsen und Nicht-Ellipsen um: es gibt weniger nicht eindeutige Nicht-Ellipsen als Ellipsen.

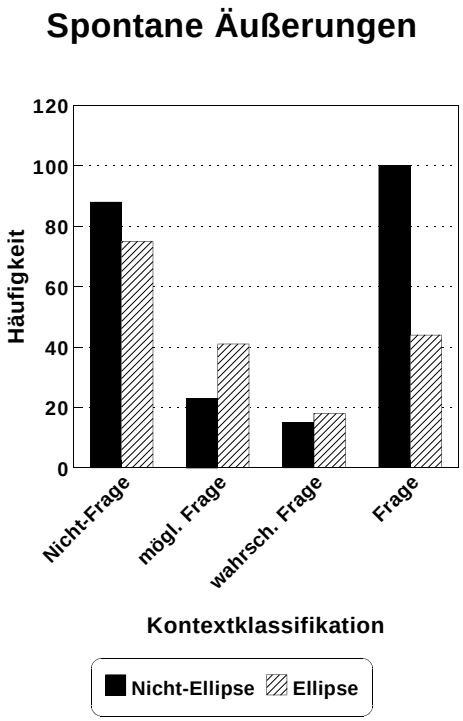

Abbildung 3

In unseren bisherigen Untersuchungen ([Bat89a, Bat91]) hat sich die Höhe des (normierten) finalen Grundfrequenzwertes ($F0$-Offset) als relevantestes prosodisches Merkmal zur Frage/-

Nicht-Frage-Unterscheidung herausgestellt. Hoher vs. tiefer $F0$-Offset entspricht grosso modo steigendem vs. fallendem Tonverlauf oder hohem vs. tiefem Grenzton. Dieses Merkmal ist eingeführt, demnach vertraut und leichter vorstellbar als etwa eine Diskriminanzfunktion, in die sechs verschiedene Merkmale eingehen. Es wird daher in der Form "$F0$-Offsetwert in Halbtönen, normiert[3] zum $F0$-Mittelwert der gesamten Äußerung", von nun an einfach "Offset" genannt, den folgenden Auswertungen zugrundegelegt. Die Normierung zum Mittelwert der Äußerung entspricht einer Normierung zum sprecherspezifischen Basiswert; sie dient dazu, unterschiedliche Stimmlagen anzugleichen. Man beachte, daß wir dieses Merkmal noch nicht optimiert haben. So ist etwa eine Normierung zur Deklinationslinie (Regressionsgeraden) vorstellbar; es ist weiter vorstellbar, daß, insbesondere bei längeren Äußerungen, eine Normierung zum $F0$-Mittelwert der Gesamtäußerung weniger günstig ist als eine zur letzten intonatorischen Phrase bzw. zur finalen syntaktischen Konstituente; vgl. dazu auch weiter unten.

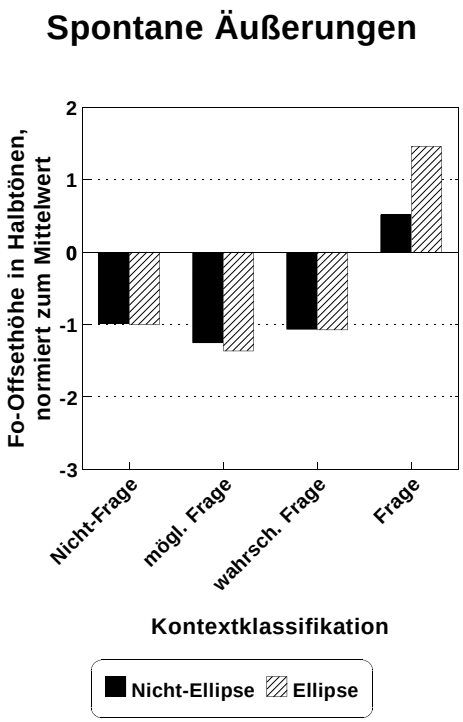

Abbildung 4

Abb. 4 zeigt für die vier Kontextklassen die Verteilung der Offset-Mittelwerte für Nicht-Ellipsen vs. Ellipsen bei SPONTAN, Abb. 5 die bei GELESEN. In beiden Registern geht eine Erhöhung der Fragehaltigkeit mit einer Erhöhung des Offsetwertes einher. Bei SPONTAN unterscheiden sich die ersten drei Klassen nicht wesentlich untereinander. Bei GELESEN erhöhen sich die Offset-Mittelwerte fast linear zur (potentiellen) Fragehaltigkeit der Äußerungen. Die anhand von Perzeptionsexperimenten entwickelte Annahme von [Bat89b], daß es Abstufungen der Fragehaltigkeit gibt, die sich auch in Abstufungen der intonatorischen Markierung widerspiegeln, scheint sich also, zumindest für GELESEN, zu bestätigen. Möglicherweise kann die Abstufung der Fragehaltigkeit bei SPONTAN durch andere nicht-linguistische, etwa situative,

[3]Der $F0$-Mittelwert wird vom $F0$-Offsetwert abgezogen.

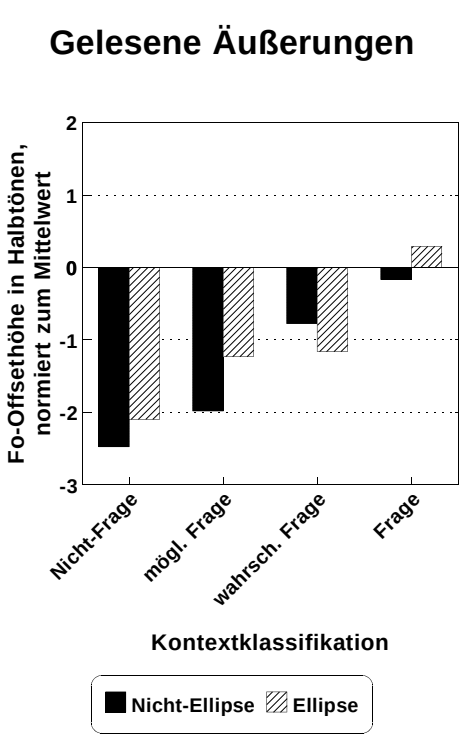

Abbildung 5

Merkmale signalisiert werden, so daß die intonatorische Markierung weniger belastet ist.

Bei der Hörerklassifikation entschieden jeweils 10 Hörer auf Frage, Aussage, Exklamativ oder Imperativ. In den Abbildungen 7 und 8 sind die drei letzten Klassen als "Nicht-Fragen" zusammengefaßt. Abb. 6 zeigt zuerst die Beziehung zwischen der Hörerklassifikation und der Kontextklassifikation für alle Äußerungen. Die Zahl der Hörer, die auf "Frage" erkannten, kann zwischen 0 und 10 liegen. Dieser Bereich ist auf den Bereich zwischen 0.0 und 1.0 abgebildet. (So bedeutet der hohe schwarze Balken ganz links, daß etwa 370 Fälle von allen Hörern und auch von der Kontextklassifikation eindeutig als Nicht-Fragen eingestuft wurden.) Das Ergebnis ist sinnvoll interpretierbar: die Nicht-Frage-Zuweisung nimmt exponentiell von links nach rechts ab, und die Frage-Zuweisung spiegelbildlich von rechts nach links. Es gibt relativ wenig uneindeutige Fälle mit Werten zwischen 0.3 und 0.7. Eine eindeutige Nicht-Frage-Zuweisung fällt leichter als eine eindeutige Frage-Zuweisung, vgl. den hohen schwarzen Balken ganz links.

In den nächsten beiden Abbildungen 7 und 8 wird die Kontextklassifikation ebenfalls den Hörerurteilen gegenübergestellt, wobei die Hörerurteile binarisiert sind: eine Äußerung gilt dann als "gehörte Frage", wenn fünf oder mehr der 10 Hörer sie als Frage klassifizierten; ansonsten gilt sie als Nicht-Frage. Es zeigt sich, daß der Unterschied zwischen SPONTAN und GELESEN vernachlässigt werden kann. Es war zu erwarten, daß die Hörer bei den beiden mittleren Kategorien breit streuen. In ca. 5% der Fälle gibt es eine Diskrepanz an den Rändern, bei den eindeutigen Kategorien Nicht-Frage bzw. Frage. Ursache dafür sind inhärente Schwierigkeiten und/oder Ambiguitäten bei den Kategorisierungen: die Äußerung *"hast nix mehr"* etwa wurde von der Partnerin mit der Frage *"hast du noch was"* beantwortet und damit kontextuell als Nicht-Frage klassifiziert, weil eine eindeutige Frage und keine Nicht-Frage folgte. Bei den

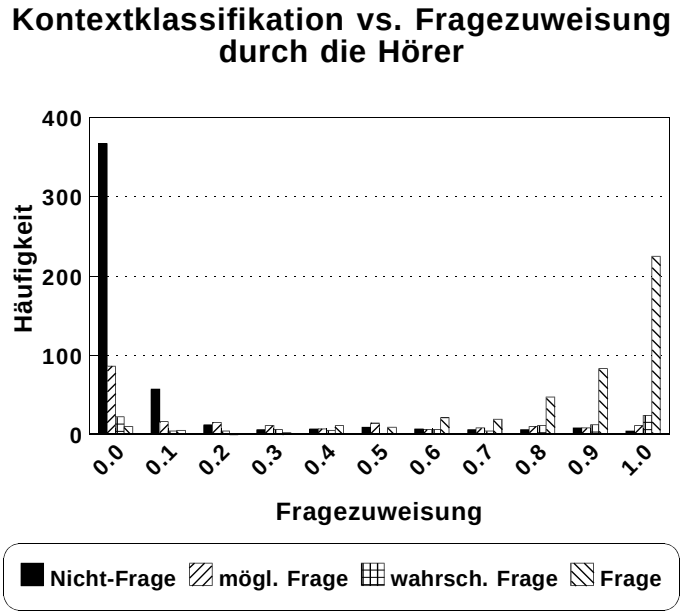

Abbildung 6

Hörerurteilen haben fünf Hörer auf Frage und fünf auf Nicht-Frage entschieden; bei der Binarisierung wurde diese Äußerung deshalb als Frage eingestuft. Beide Interpretationen (Frage bzw. Aussage mit fehlendem Vorfeld) sind möglich. Der Offset hat zwar einen positiven Wert, disambiguiert aber offensichtlich auch nicht eindeutig.

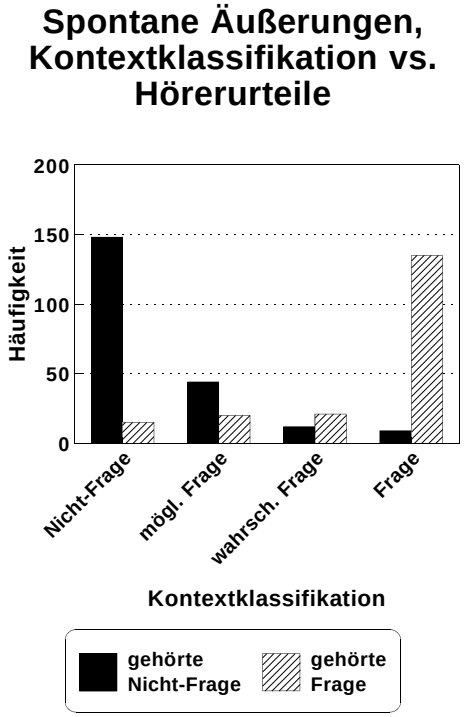

Abbildung 7

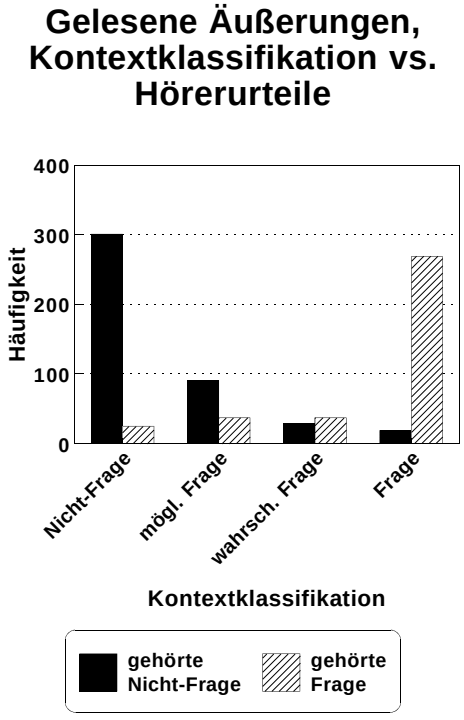

Abbildung 8

In den Abbildungen 9 und 10 sind die Ergebnisse der Hörtests aufgetragen, getrennt nach SPONTAN/GELESEN (man beachte, daß genau doppelt soviel gelesene wie spontane Äußerungen vorlagen), und aufgeschlüsselt nach Nicht-Ellipsen/Ellipsen. Da im Augenblick nur die Frage/Nicht-Frage-Unterscheidung interessiert, ist wieder nur die "Frage"-Klassifikation angegeben, abgebildet auf den Bereich zwischen 0 und 1. Da pro Äußerung immer je 10 Versuchspersonen eine Antwort abgaben, bedeuten z.B. in Abb. 9 die beiden linkesten Balken bei 0.0, daß in gut 80 Fällen der Nicht-Ellipsen und in knapp 80 Fällen der Ellipsen alle zehn Hörer auf Frage entschieden. Zwischen SPONTAN und GELESEN gibt es keine großen Unterschiede, wohl aber zwischen Nicht-Ellipsen und Ellipsen: Ellipsen, die nicht eindeutig (0.0) als Nicht-Fragen klassifizierbar sind, können offensichtlich weniger eindeutig klassifiziert werden als Nicht-Ellipsen. Dies zeigt sich insbesondere beim Wert 1.0 (100%ige Frage-Zuweisung).

Die Abbildungen 11 und 12 zeigen für SPONTAN und GELESEN den Offset-Mittelwert pro Anzahl Versuchspersonen (analog zu den Abbildungen 9 und 10), wieder abgebildet auf den Bereich zwischen 0 und 1, aufgegliedert nach Nicht-Ellipsen und Ellipsen. Die beiden linken Balken in Abb. 11 bedeuten also, daß der Mittelwert des Offsets in den Fällen, bei denen alle Hörer auf "Nicht-Frage" entschieden, knapp 2 Halbtöne unter dem Äußerungsmittelwert lag. Im mittleren Wertebereich von 0.3 bis 0.7 gibt es relativ wenig Fälle pro Klasse, vgl. Abb. 9 und 10. Einzelne Extremwerte haben also einen verfälschenden Einfluß auf den Mittelwert. Wir haben deshalb alle Fälle mit Werten von 0.3 bis 0.7 zusammengefaßt und auf den mittleren Wert 0.5 projiziert. Global kann man sagen, daß umso mehr Hörer auf Frage entschieden, je höher der Offset war. (Zur Frage nach eventuellen systematischen Unterschieden zwischen einzelnen Satztypen, die durch die Mittelwertbildung überdeckt werden, vgl. den nächsten Abschnitt.) Wieder ist – wie bei den Abb. 4 und 5 – das Bild bei GELESEN eindeutiger und regelmäßiger als bei SPONTAN, und bei den Ellipsen deutlicher als bei den Nicht-Ellipsen; man beachte besonders den fast linearen Anstieg in Abb. 12 bei den gelesenen Ellipsen. Offensichtlich wird die Fragezuweisung bei Ellipsen mangels anderer indizierender Merkmale stärker über die Intonation und damit über den Offset gesteuert als bei Nicht-Ellipsen und offensichtlich setzen die Sprecher die Intonation bei GELESEN systematischer ein als bei SPONTAN; dort wiederum sind die Offsetwerte höher als bei GELESEN, vgl. [BJKN92].

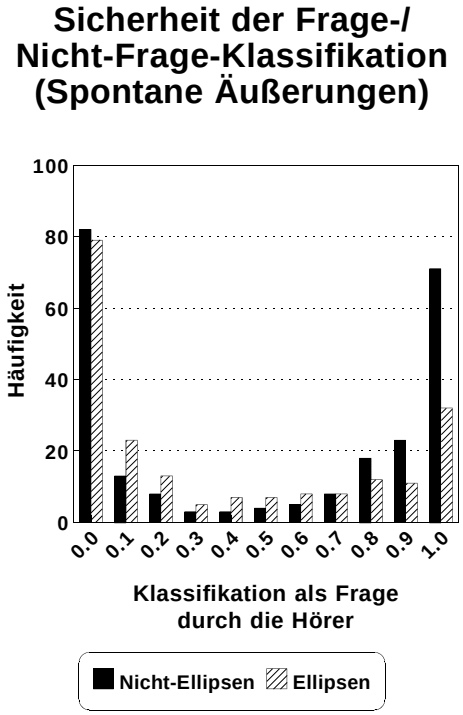

Abbildung 9

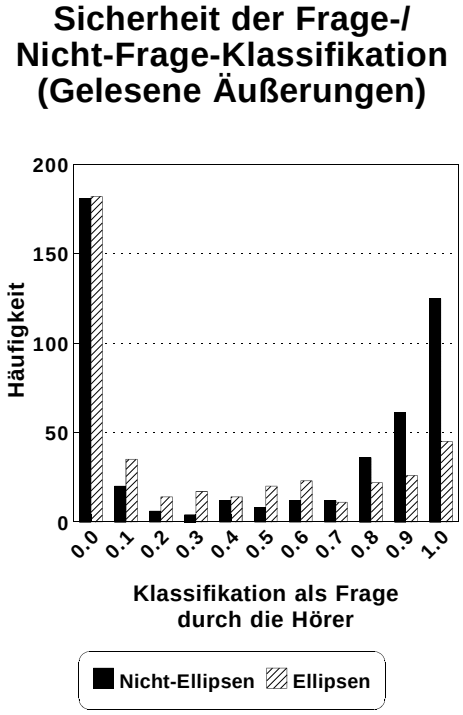

Abbildung 10

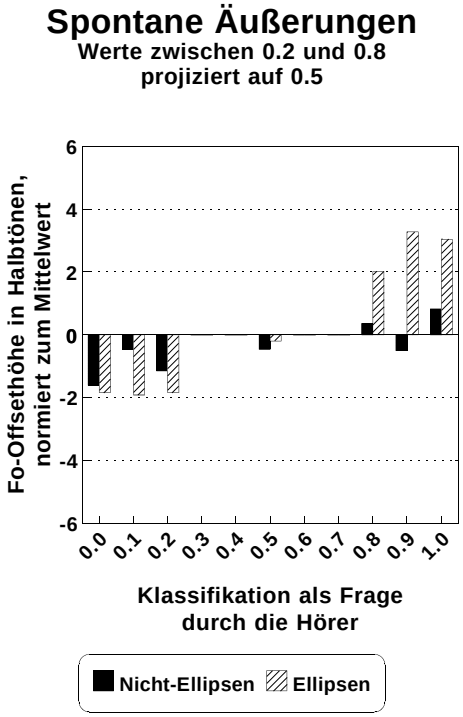

Abbildung 11

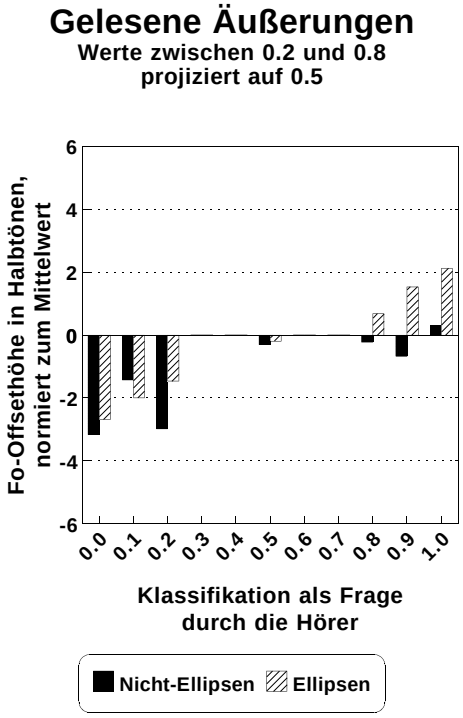

Abbildung 12

5 Offsethöhe bei unterschiedlichen Satztypen

[DZ90] berichten für das amerikanische Englisch, daß in einem spontansprachlichen Korpus aus der Mensch-Maschine-Kommunikation 36% der Entscheidungsfragen ("YES-NO-Questions") und über 90% der Ergänzungsfragen ("WH-Questions") einen tiefen Grenzton (und damit normalerweise einen tiefen Offset) aufweisen. Entsprechende Angaben fehlen bisher für die deutsche Spontansprache. Bei unserem Material ist eine analoge deskriptive Statistik nur bei den Satztypen sinnvoll, die relativ oft vorkommen, d.h. bei den Ergänzungsfragen, den Entscheidungsfragen, den Aussagen und den unterschiedlichen Imperativen, zusammengefaßt als Typklasse Imperativ. Die nächsten Abbildungen 13 bis 17 zeigen daher in einem Histogramm (Bereich -10 bis +10 Halbtöne über dem Mittelwert) die Verteilung der Offsethöhe für alle Äußerungen sowie für die erwähnten Satztypen. Die Grenze der Diskriminanzfunktion (und damit die Grenze zwischen Frage-/Nicht-Fragezuweisung bei der automatischen Klassifizierung, vgl. unten) liegt etwas über 0 Halbtönen. Die schwach bimodale Verteilung bei allen Äußerungen ist erwartet und entspricht der in [Bat91], die an einem elizitierten Korpus ermittelt wurde. Bei den beiden Fragetypen ist die Verteilung deutlich bimodal. Aussagen sind zwar wie Imperative unimodal, aber stark rechtsschief. Mögliche Ursachen sind schon in [Bat91] angegeben: Wenn bei einer Aufzählung *"Eins, zwei, drei, vier, fünf, sechs, sieben, acht."* etwa der Satzakzent auf *"acht"* liegt, so ist der Offset natürlich klar über dem Mittelwert, ohne daß aber damit eine Frage indiziert wurde.

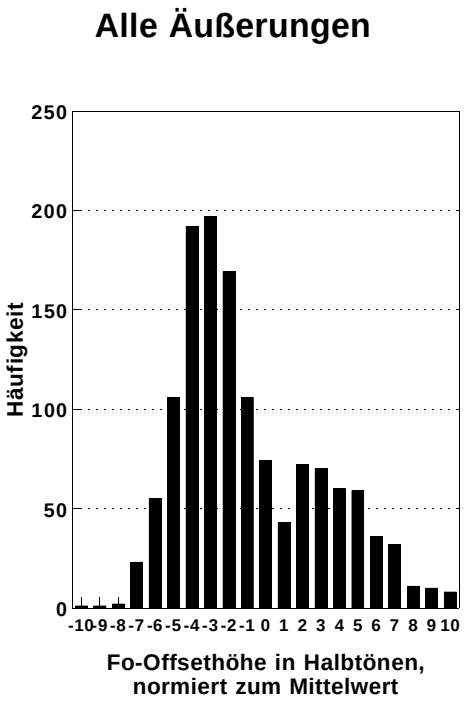

Abbildung 13

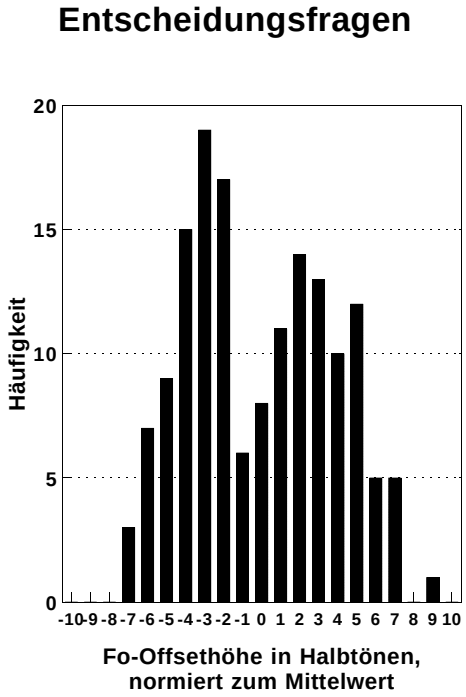

Abbildung 14

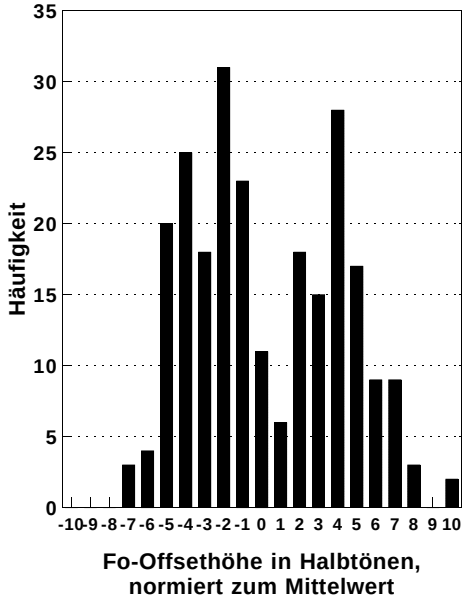

Abbildung 15

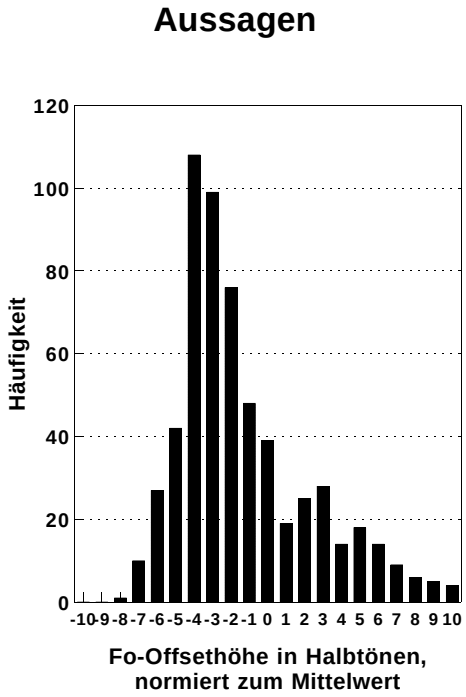

Abbildung 16

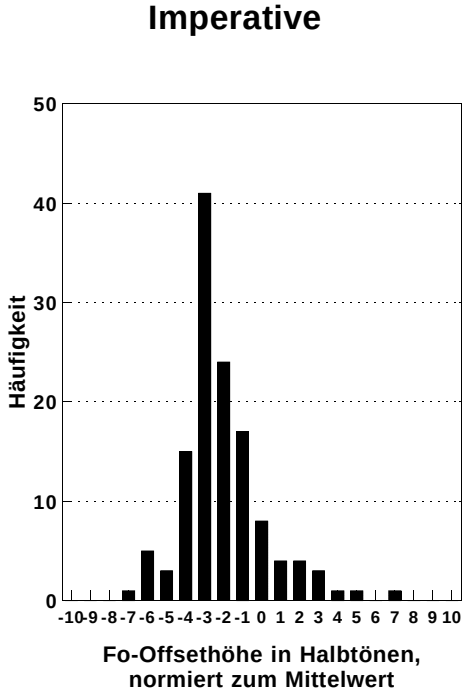

Abbildung 17

Vergleichbar mit den Ergebnissen von [DZ90] sind die Resultate der folgenden vier Klassen: Ergänzungsfragen (WH-Questions), Entscheidungsfragen (YES-NO-Questions), Ausagen (Statements) und Imperative (Commands). Bei [DZ90] wurde die Zuweisung zu Hochton bzw. Tiefton durch Experten perzeptiv vorgenommen. Eine automatische Berechnung und Dichotomisierung mit dem Merkmal "$F0$-Mittelwert der letzten Silbe, normiert zum $F0$-Mittelwert der Gesamtäußerung" stimmte bei [DZ90] mit den Perzeptionsurteilen zu 90 % überein. Wir definieren für den Vergleich einfach positive Offsetwerte als hohe und negative als tiefe Grenztöne[4]. Abb. 18 zeigt für SPONTAN sowie für das englische Material den Prozentsatz der Fälle mit tiefem Grenzton, wobei unser Material noch aufgeteilt ist nach nicht-elliptisch und elliptisch; nachdem [DZ90] auch eine Klasse "Fragment" anführen, die allerdings nicht deckungsgleich mit unseren Ellipsen sein dürfte, wird es sich bei den verglichenen englischen Fällen um nicht-elliptische Äußerungen handeln. Abb. 19 zeigt analog zu Abb. 18 den Vergleich mit GELESEN. Durchgehend läßt sich feststellen, daß es bei SPONTAN etwa um 10 % weniger tiefe Offsetwerte gibt als bei GELESEN. Das mag damit zusammenhängen, daß bei SPONTAN der Offset im Schnitt höher ist als bei GELESEN, vgl. [BJKN92]. Deutlich ist der Unterschied zwischen den beiden Sprachen bei den beiden Fragetypen: unser Material ist bei den Nicht-Ellipsen grob gleichverteilt, bei den Ellipsen überwiegen die Fälle mit hohem Grenzton. Die Imperative unterscheiden sich kaum. Die nicht ganz erwartete hohe Anzahl von Aussagen mit hohem Offset zeigte sich schon in Abb. 16 bei den Werten > 0, vgl. die Diskussion im letzten Abschnitt. Es verwundert auf den ersten Blick, daß die Aussagen, insbesondere die nicht-elliptischen, nicht mehr Fälle mit tiefem Grenzton aufweisen. Wir vermuten, daß bei dem Unterschied zwischen Ellipsen und Nicht-Ellipsen auch die Gesamtdauer der Äußerungen eine Rolle spielt. Tab. 2 zeigt die Dauermittelwerte und die Zahl der Fälle der mit [DZ90] verglichenen Satztypen. Bei den Aussagen sind die Nicht-Ellipsen im Schnitt doppelt so lang wie die Ellipsen; vgl. dazu die Anmerkungen zur fehlenden Optimierung weiter oben und im nächsten Teil. Da die Zahl der Fälle bei den elliptischen Entscheidungsfragen sehr klein ist, ist der entsprechende Wert in Abb. 18 und Abb. 19 mit Vorsicht zu interpretieren.

Das Material von [DZ90] unterscheidet sich von unserem in mehreren Aspekten: Mensch-Maschine- vs. Mensch-Mensch-Kommunikation, 89 Sprecher vs. vier Sprecher, perzeptive Klassifizierung nach "Hoch/Tief" vs. (suboptimale) automatische Ableitung, etc. Alle diese Faktoren können eine Rolle spielen; die Annahme, daß es sich einfach um einen Unterschied zwischen den beiden Sprachen Englisch und Deutsch handelt, bleibt aber bis zum Beweis des Gegenteils die plausibelste.

[4]Diese einfache Definition wird allerdings sicher nicht immer deckungsgleich sein mit einer perzeptiven Evaluierung. Auch bei [DZ90] zeigte sich ja eine Abweichung von 10 %.

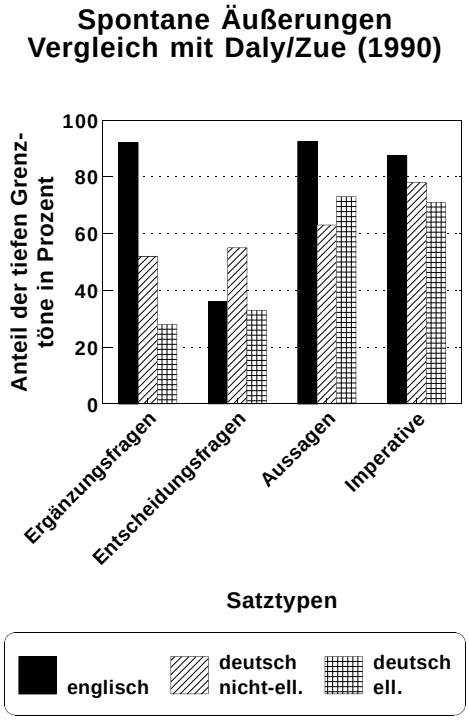

Abbildung 18

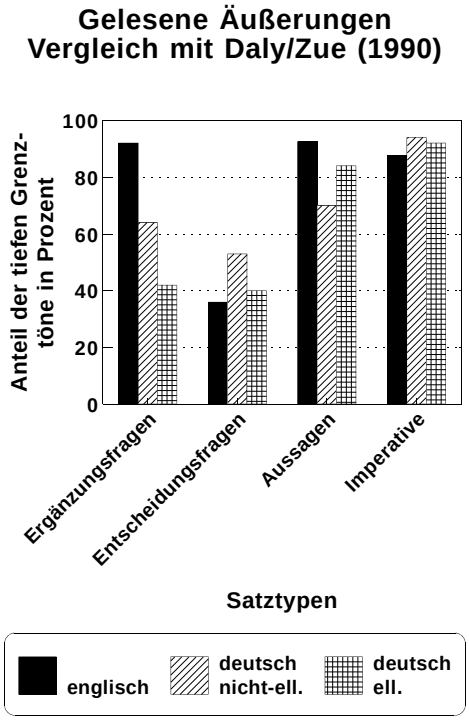

Abbildung 19

		Gesamtdauer-Mittelwert in msec.	Zahl der Fälle
Ergänzungsfragen	Nicht-Ellipsen	697.4	168
	Ellipsen	419.6	75
Entscheidungsfragen	Nicht-Ellipsen	586.9	147
	Ellipsen	610.8	8
Aussagen	Nicht-Ellipsen	951.9	245
	Ellipsen	437.2	348
Imperative	Nicht-Ellipsen	387.4	108
	Ellipsen	584.0	20

Tabelle 2: Dauermittelwerte und Zahl der Fälle der mit [DZ90] verglichenen Satztypen.

6 Automatische Klassifikation

Abb. 20 zeigt für alle 1329 Fälle das Ergebnis von Diskriminanzanalysen mit Lern=Test und mit unterschiedlichen Merkmalen als Einzelprädiktoren (univariate Analyse) und allen Merkmalen zusammen (multivariate Analyse). Es handelt sich dabei um Onset, Offset, Maximum, Minimum (alle in Halbtönen normiert zum Mittelwert), sowie um den Range (Betragsdifferenz von Maximum und Minimum), die Streuung und den Regressionskoeffizienten (die letzten drei Werte ebenfalls in Halbtönen).

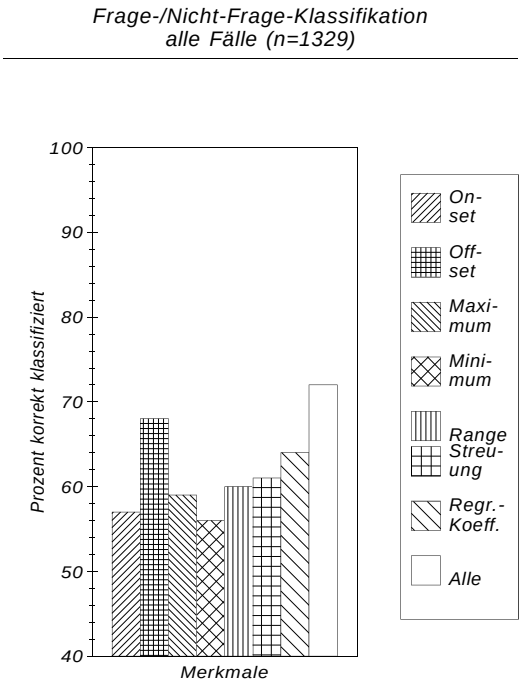

Abbildung 20

Wie schon weiter oben erwähnt, geschah die Frage/Nicht-Frage-Klassifikation auf der Grundlage der Hörerklassifikation: Wenn weniger als fünf Hörer auf Frage entscheiden, wurde die Äußerung als Nicht-Frage geführt, in allen anderen Fällen als Frage. Wie bei unseren anderen Untersuchungen an elizitiertem Material (vgl. [Bat89a, Bat91]) erweist sich der Offset als relevantestes einzelnes Merkmal; der damit positiv korrelierende Regressionskoeffizient (vgl. [BJKN92]) ist etwas schlechter. Alle Merkmale zusammen ergeben allerdings eine um etwa 5% bessere Klassifikation. In den nächsten vier Abbildungen 21 bis 24 sind die analogen Ergebnisse für SPONTAN, GELESEN, ELLIPSEN und NICHT-ELLIPSEN aufgetragen. Die Unterschiede zwischen SPONTAN und GELESEN sind unauffällig, Ellipsen werden besser klassifiziert als Nicht-Ellipsen; dieses Ergebnis war erwartet, da sich schon in der deskriptiven Statistik Ellipsen von Nicht-Ellipsen unterschieden. Die Klassifikationsgüte ist deutlich schlechter als bei dem stark kontrollierten und eingeschränkten Korpus von [Bat89a], wo mit vergleichbaren Merkmalen 97% erreicht werden konnten, und um ca. 10% schlechter als in [Bat91], wo ebenfalls ein elizitiertes, aber in sich unterschiedlicheres Material (intonatorische Minimalpaare) untersucht wurde. Es scheint sich hierbei doch eher um einen Effekt zu handeln, der auf das konstruier-

te Material zurückzuführen ist, und nicht auf den Unterschied GELESEN/SPONTAN, der in unserem jetztigen Material unauffällig bleibt. Durchgehend und deutlich ist aber der Unterschied zwischen den bisher noch nicht untersuchten Ellipsen und den Nicht-Ellipsen. Eine für beide Konstellationen zutreffende Annahme ist, daß Sprecher offensichtlich bei möglichen Ambiguitäten dazu tendieren, die Intonation stärker zur Disambiguierung einzusetzen. Konstruierte intonatorische Minimalpaare sind potentiell ebenso ambig wie elliptische Strukturen.

Die Abbildung 25 zeigt, daß zwischen den einzelnen Sprechern zwar sichtbare, aber nicht zu große Unterschiede bestehen.

Die fehlende Optimierung der Merkmale dürfte eine etwas zu schlechte, die Gleichsetzung von Lern- und Teststichprobe eine etwas zu gute Klassifikation zur Folge haben. Neben anderen Normierungen sollte in Zukunft auf alle Fälle untersucht werden, ob (insbesondere bei längeren Äußerungen) eine Beschränkung auf einen finalen Bereich bei der Mittelwertbildung und der anschließenden Offsetnormierung nicht bessere Ergebnisse liefert; eine Alternative dazu ist die Bildung des Quotienten des Mittelwertes der letzten Silbe – nicht eines punktuellen finalen Wertes – zum Äußerungsmittelwert; vgl. [DZ90]. Bei Äußerungen mit finalem Satzakzent kann man versuchen, den dann oft vorliegenden fragetypischen finalen Tonverlauf (sog. “Tiefton” im Ton-Sequenzansatz) von einem aussagetypischen finalen Tonverlauf (“Hochton”) abzugrenzen. Eine entscheidende Verbesserung bei einer automatischen Klassifikation ist aber wohl am ehesten in Verbindung mit anderen Analyseebenen (Syntax, lexikalische Füllung, Dialogwissen) zu erwarten, da die mit tiefem Offset produzierten Entscheidungs- und Ergänzungsfragen wohl kaum mit anderen intonatorischen Merkmalen richtig klassifizierbar sind. Eine andere Möglichkeit dürfte sein, unklare Fälle auszusondern: unklar können z.B. Fälle mit wenig ausgeprägtem Offset sein, oder Fälle mit finalem Satzakzent und damit der potentiellen Vermischung von Modus- und Fokusindizierung.

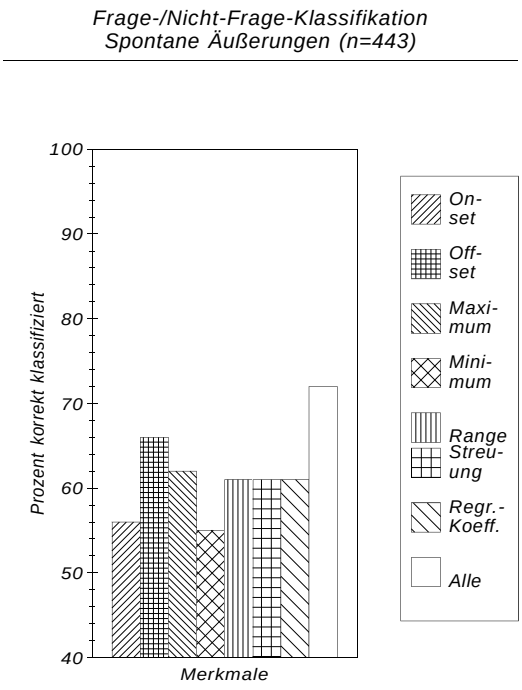

Abbildung 21

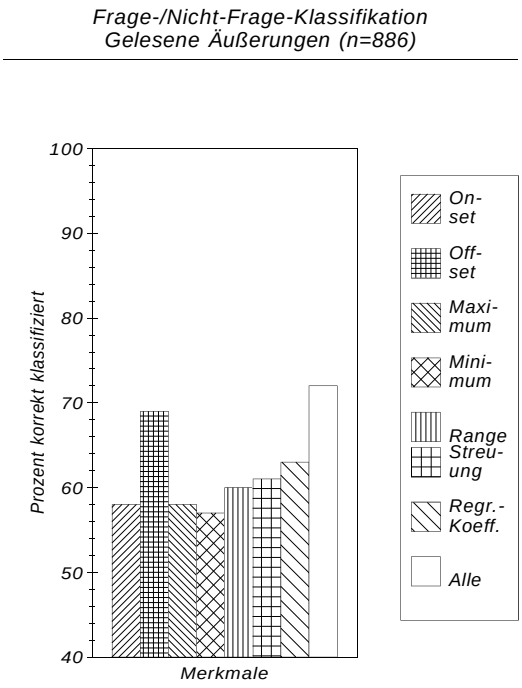

Abbildung 22

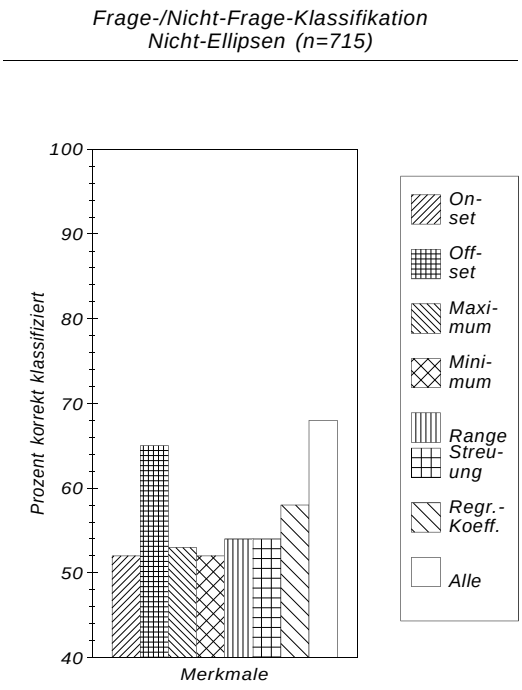

Abbildung 23

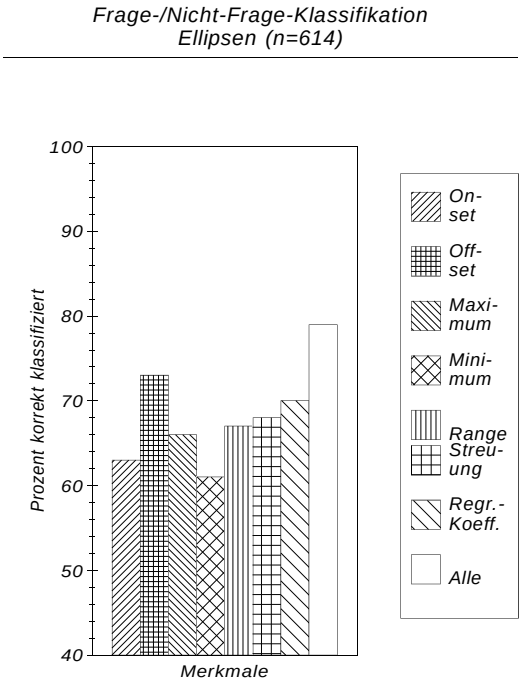

Abbildung 24

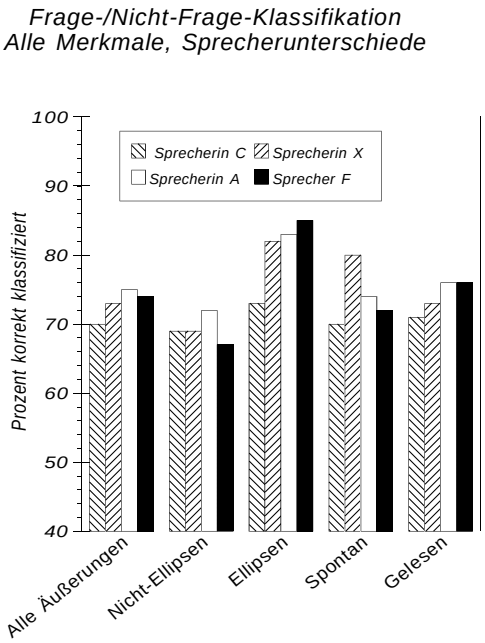

Abbildung 25

7 Zusammenfassung

Es hat sich gezeigt, daß eine Abstufung der Fragehaltigkeit einhergeht mit einer Abstufung der indizierenden intonatorischen Markierung; die Frage-/Nicht-Frage-Dichotomie sollte deshalb nicht als eine strikt binäre aufgefaßt werden. Der Unterschied zwischen spontanem und gelesenem Material war in der Ausprägung der Kennwerte nicht sehr auffällig, auch wenn sich das gelesene Material systematischer verhält. Hingegegen gab es einen durchgehenden Unterschied zwischen nicht-elliptischen und elliptischen Äußerungen: bei den letzteren wird offensichtlich die Intonation mangels anderer Disambiguierungsmöglichkeiten stärker eingesetzt. Grundsätzlich bestätigten sich ansonsten die an elizitiertem, gelesenem Material gewonnenen Ergebnisse der automatischen Klassifizierung.

Bei Entscheidungs- und Ergänzungsfragen ist eine Tendenz zur Gleichverteilung von hohem und tiefem Offset festzustellen, wobei die intonatorische Fragemarkierung durch einen hohen Offset zumindest zum Teil funktional eingesetzt werden dürfte, z.B. dann, wenn der Hörer stärker zu einer antwortenden Reaktion verpflichtet werden soll. Andere mögliche Faktoren, etwa sprecher- oder geschlechtsspezifische Strategien, konnten mit den bisherigen Auswertungen und der zugrundeliegenden Materialbasis nicht erfaßt werden.

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie (Förderkennzeichen 01IV102F4 und 01IV102H0) und der Deutschen Forschungsgemeinschaft (Al 173/4) gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Literatur

[Alt87] H. Altmann: *Zur Problematik der Konstitution von Satzmodi als Formtypen*, in J. Meibauer (Hrsg.): *Satzmodus zwischen Grammatik und Pragmatik*, Niemeyer Verlag, Tübingen, 1987, S. 22–56.

[Bat89a] A. Batliner: *Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen*, in H. Altmann, A. Batliner, W. Oppenrieder (Hrsg.): *Zur Intonation von Modus und Fokus im Deutschen*, Max Niemeyer Verlag, Tübingen, 1989, S. 21–70.

[Bat89b] A. Batliner: *Wieviel Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorien*, in H. Altmann, A. Batliner, W. Oppenrieder (Hrsg.): *Zur Intonation von Modus und Fokus im Deutschen*, Max Niemeyer Verlag, Tübingen, 1989, S. 111–162.

[Bat91] A. Batliner: *Ein einfaches Modell der Frageintonation und seine Folgen*, in E. Klein, F. P. Duteil, K. Wagner (Hrsg.): *Betriebslinguistik und Linguistikbetrieb*, Niemeyer, Tübingen, 1991, S. 147–160.

[BJKN92] A. Batliner, B. Johne, A. Kießling, E. Nöth: *Zur prosodischen Kennzeichnung von spontaner und gelesener Sprache*, in G. Görz (Hrsg.): *Proc. Konferenz Verarbeitung Natürlicher Sprache: KONVENS 92, "Informatik aktuell"*, Springer Verlag, Berlin, Heidelberg, 1992, S. 29–38.

[BKKN91] A. Batliner, A. Kießling, R. Kompe, E. Nöth: *"Irregularitäten" spontaner Sprache und ihre Verarbeitung mit automatischen Grundfrequenzverfahren*, in *Proc. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik – DAGA*, Bd. B, DPG–GmbH, Bad Honnef, 1991, S. 993–996.

[DZ90] N. Daly, V. Zue: *Acoustic, Perceptual, and Linguistic Analyses of Intonation Contours in Human/Machine Dialogues*, in *Int. Conf. on Spoken Language Processing*, Kobe, 1990, S. 497–500.

[Hes91] W. Hess: *Persönliche Mitteilung*, 1991, Institut für Kommunikationsforschung und Phonetik, Universität Bonn.

[KKN*92] A. Kießling, R. Kompe, H. Niemann, E. Nöth, A. Batliner: *DP-Based Determination of F0 Contours From Speech Signals*, in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Bd. 2, San Francisco, 1992, S. II–17–II–20.

[Ree89] H. Reetz: *A Fast Expert Program for Pitch Extraction*, in *Proc. European Conf. on Speech Communication and Technology*, Bd. 2, Paris, 1989, S. 476–479.