DIE MODELLIERUNG VON LAUTDAUERVARIATIONEN IM ÖSTERREICHISCHEN DEUTSCH

Hannes Pirker ***Friedrich Neubarth***

Österreichisches Forschungsinstitut für Artificial Intelligence (ÖFAI) *
Schottengasse 3, A–1010 Wien
Email: hannes@ai.univie.ac.at

1 EINLEITUNG

Das Projekt "SpeeDurCont" (*Spee*ch *Dur*ation in *Cont*ext-to-Speech) widmet sich der Untersuchung von Lautdauervariationen im "Österreichischen Deutsch".

Es zielt darauf ab, die zahlreichen Einflüsse auf die Lautdauer zu quantifizieren. Aus praktischer Sicht sollen die so gewonnenen Dauermodelle die prosodische Qualität eines Sprachsynthetisators [Rank & Pirker 1998] erhöhen. Die Ergebnisse sollen aber auch das theoretische Verständnis der Wirkung einzelner Faktoren vertiefen.

Als Methode werden in SpeeDurCont Maschinelle Lernverfahren verwendet, die auf einer eigens erstellten Sprachdatensammlung angewendet werden.

Dieser Beitrag stellt den Aufbau des verwendeten Korpus dar und diskutiert die Anforderungen an mögliche Kodierungsstrategien für die Datensammlung.

2 PROBLEMSTELLUNG

Die Dauer eines Lautes in einer Äußerung wird von zahlreichen Faktoren bestimmt: Intrinsische Lautdauer und segmentaler Kontext (die Identität oder Phonemklassezugehörigkeit von Nachbarlauten), metrische Einflüsse (z.B. je mehr Silben ein Wort umfaßt desto kürzer die Lautdauer), Akzentuierung, Prominenz, Position in einer prosodischen Konstituente (z.B. finale Dehnung) uam.

Daß diese Faktoren jeweils die Lautdauer beeinflussen kann als gesichert gelten. Für die Quantifizierung der kombinierten Effekte stellt sich jedoch das Problem, daß die Anzahl der möglichen Faktor-Wert-Kombinationen in einem groben Mißverhältnis zur Anzahl der Beispieldaten in den zur Verfügung stehenden Korpora steht ("data sparsity").[1]

Das prinzipielle Problem kann nicht einfach durch eine Vergrößerung der Datenbasis gelöst werden, da dieser aufgrund des für die Datenaufbereitung nötigen manuellen Aufwands enge Grenzen gesetzt sind.

Daher muß neben der Verwendung eines möglichst geeigneten Lernverfahrens darauf abgezielt werden, eine effiziente Vorauswahl von erfolgversprechenden Faktoren zu treffen um die Anzahl möglicher Klassen zu reduzieren.

Nicht zuletzt weil also die Klassenauswahl nicht von vornherein festgelegt sein kann, und erst durch prospektive Auswertungen erfolgt, wird in einem ersten Schritt eine möglichst umfassende Datenbasis aufgebaut. Dabei wird großes Augenmerk auf die Repräsentationsform gelegt. Sie soll ermöglichen, die für Sprachkorpora typische Mischung aus heterogenen, sequentiellen und hierarchisch organisierten Informationen in flexibler und transparenter Weise zu kodieren.

3 SPRACHKORPUS

Der für die Untersuchung verwendetet Korpus umfaßt 50.000 Laute eines einzelnen Sprechers, und ist somit vom Umfang her um einiges größer als bislang in diesem Zusammenhang verwendeten Korpora des Deutschen. Die Sprachdaten wurden in einem schallarmen Raum mittels DAT aufgenommen, für die aktuelle Annotationsarbeit wurden Tondateien mit einer auf 44.1kHz und 16kHz reduzierter Abtastrate verwendet. Die Aufnahmen enthalten auf einem zweiten Kanal ein Laryngographen-Signal.

Der Korpus enthält die folgenden gelesenen Texte:

250	Frage/Antwort-Paare (FA-Sätze)
300	Einzelsätze ("Marburg-Sätze",...)
24	Kurze Texte (v.a. Zeitungsartikel)

Ein Teil der Texte wurde auch im bekannten PhonDat-Korpus [Kohler 1994] verwendet. Durch diese Wahl bietet sich die Möglichkeit bei Bedarf diesen Multi-Sprecher Korpus für vergleichende Untersuchungen zu verwenden.

Eine Besonderheit stellen jedoch die FA-Sätze dar. Sie wurden mit dem Ziel zusammengestellt, in einem Spezialkorpus Syntax, Phonologie und insbesondere die Informationsstruktur systematisch zu variieren, und so deren Einfluß auf die Prosodie zu untersuchen.

Der FA-Korpus besteht aus 250 Paaren von Fragen und Antworten, wobei die Fragen dazu dienen die Informationsstruktur oder Fokusbedingungen der Antwortsätze festzulegen. Nur die Antworten werden ausgewertet.

Beispiel (A1/A2) zeigt, wie sich eine Änderung von weitem zu engem Fokus auf die Position der Akzente (durch Kapitalisierung verdeutlicht) und auch die prosodische Phrasierung (hier mit runden Klammern

* Das ÖFAI wird vom *Österreichischen Bundesministerium für Wissenschaft und Verkehr* unterstützt. Die gegenständliche Arbeit wurde zudem vom *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)* im Rahmen des Projekts P13224 gefördert.

[1] Beispielsweise finden sich bei [Riedi 1995] in einer Datenbasis von 21.500 Lauten 15.850 verschiedene Klassen, ein durchaus typisches Zahlenverhältnis.

markiert) auswirkt:

F1: *Was ist los?*
A1: (Der FREUND) (verspricht der Direktorin zu ARBEITEN) (und das BÜRO zu putzen)
F2: *WEM verspricht der Freund zu arbeiten?*
A2: (Der Freund verspricht der DIREKTORIN zu arbeiten) (und das BÜRO zu putzen)
F3: *Was ist los?*
A3: (Der FREUND verspricht) (die DIREKTORIN zu entlasten) (und das BÜRO zu putzen)

Die Variation auf syntaktischer Ebene wird im Beispiel (A1/A3) demonstriert: durch den Wechsel in der Transitivität des eingebetteten Verbs ("entlasten") ändert sich wiederum die prosodische Phrasierung und parallel dazu die Akzentpositionen.

Außerdem werden die metrischen Verhältnisse in den Konstituenten variiert, indem Wörter mit unterschiedlicher Silbenanzahl in derselben syntaktischen Position verwendet werden (z.B. "Udo" statt "die Direktorin").

4 ANNOTATIONEN

Die Annotierenung des Korpus, die noch nicht abgeschlossen ist, umfaßt folgende Elemente:
Lautsegmentierung: Die FA-Sätze sind manuell eng phonetisch transkribiert. Dabei wurden Abweichungen von einer postulierten kanonischen Transkription analog zu PhonDat speziell notiert. In der Folge werden diese Daten zum Trainieren eines automatischen Aligners [Stöber & Hess 1998] verwendet, mit dessen Ergebnissen die Segmentierung der restlichen Korpusteile unterstützt wird.
Prosodische Hierarchie wird weitgehend automatisch auf der Basis der Lautsegmentierung und eines Lexikons erzeugt. Es werden so die Grenzen der Phoneme, Silbenkonstituenten und Wörter ermittelt.
Intonation und Prosodische Phrasierung wird manuell unter der Verwendung der G-ToBI Konventionen [Grice et al. 96] annotiert. Zusätzlich soll diese durch eine phonetische Parametrisierung der Grundfrequenzkontur – beispielsweise durch die Verwendung von cos^2 -Splines [Portele et al. 1995] – erfolgen .
Perzipierte Prominenzen werden jeder Silbe auf einer 5-stufige Skala zugewiesen.
Syntaktische Information bleibt i.A. auf die Wortklasse beschränkt. Für den FA-Subkorpus ist die syntaktische Phrasenstruktur voll spezifiziert.
Informationsstruktur (weiter vs. enger Fokus) ist für die FA-Sätze – wie oben gezeigt – anhand der Fragestruktur definiert und kann daher für diesen Teil des Korpus ebenfalls ausgewertet werden.

5 REPRÄSENTATION

Wie die Aufzählung im vorigen Abschnitt zeigt, weist der SpeeDurCont-Korpus die für Sprachkorpora typische heterogene Struktur auf: Die Informationen stammen aus unterschiedlichen Quellen und die verschiedenen Annotations- und Analysewerkzeuge sorgen üblicherweise für die Multiplikation von Darstellungsformen und Dateiformaten.

Die Forderungen an eine Repräsentationsform für diese Daten sind nicht einfach zu erfüllen. Sie soll einerseits eine möglichst gute Wartbarkeit gewährleisten, was insbesondere bei einer noch im Aufbau befindliche Datenbasis von Bedeutung ist. Sie soll aber auch komplexe Abfragen (queries) effizient und flexibel unterstützen.

In unserem Projekt wird momentan an der Entwicklung eines Repräsentationsschemas gearbeitet, das sich am Konzept der Annotationsgraphen [Bird & Liberman 1999] aber auch am Emu-Modell [Cassidy & Harrington 1996] orientiert. Annotationsgraphen stellen ein interessantes formales Modell zur Kodierung von Sprachdaten dar, bei dem die eigentliche technische Realisierung nicht festgelegt ist.

Da in diesem Modell Zusammenhänge zwischen Beschreibungsebenen durch symbolische Relationen (Knoten im Annotationsgraph) und nicht über die Identität von Zeitmarken ausgedrückt werden, bietet es sich zur Kodierung eines möglichst redundanzfreien Basisformats an, auf dem allfällige Änderungen durchgeführt werden. Bei Bedarf können daraus ausspezifizierte Hilfsformate erzeugt werden, falls solche für die Effizienzsteigerung bei Abfragen oder für die Bearbeitung der Daten benötigt werden.

Da das zu benutzende Maschinelle Lernprogramm auf Prolog basiert, müssen schlußendlich die Daten in Form von Prologtermen vorliegen. Da sich Prolog zudem besonders gut für die Kodierung und Auswertung relationaler Repräsentationen eignet, bietet es sich als Basis für die Implementierung des Graphenmodells an.

Literatur

[Bird & Liberman 1999] Bird S., Liberman M.: A Formal Framework for Linguistic Annotation, Department of Computer and Information Science, Univ. of Pennsylvania, Philadelphia, Tech Report MS-CIS-99-01, 1999.

[Cassidy & Harrington 1996] Cassidy S., Harrington J.: EMU: an Enhanced Hierarchical Speech Data Management System, Proc. of 6th Australian Int.Conf. on Speech Science and Technology (SST-96), Adelaide, 1996.

[Grice et al. 96] Grice M., Reyelt M., Benzmüller R., Mayer J., Batliner A.: Consistency in Transcription and Labelling of German Intonation with GToBI, Proc. of ICSLP'96, Philadelphia, pp.1716-19, 1996.

[Kohler 1994] Kohler K.: Lexica of the Kiel PHONDAT Corpus, Read Speech, Inst. f. Phonetik und Digitale Sprachverarbeitung, Univ. Kiel, Vol.I, Arbeitsberichte Nr.27, 1994.

[Portele et al. 1995] Portele T., Krämer J., Heuft B., Sonntag G.: Parametrisierung von Grundfrequenzkonturen, DAGA-95, Saarbrücken, 1995.

[Rank & Pirker 1998] Rank E., Pirker H.: VieCtoS—Speech Synthesizer, Technical Overview, Österreichisches Forschungsinstitut für Artificial Intelligence, Wien, TR-98 -13, 1998.

[Riedi 1995] Riedi M.: A Neural-Network-Based Model of Segmental Duration for Speech Synthesis, Proc. Eurospeech-95, Madrid, Spain, Vol.1, pp.599-602, 1995.

[Stöber & Hess 1998] Stöber K., Hess W.: Additional Use of Phoneme Duration Hypotheses in Automatic Speech Segmentation, Proc. of ICSLP'98, Sydney, Australia, 1998.