Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Prosodische Etikettierung und Segmentierung deutscher Spontansprache

Michael Lehning

TU Braunschweig

Vm

Report 160
13. August 1996

13. August 1996

Michael Lehning

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
38092 Braunschweig

Tel.: (0531) 391 - 2476
Fax: (0531) 391 - 8218
e-mail: lehning@ifn.ing.tu-bs.de

Gehört zum Antragsabschnitt: 14.3 Werkzeuge zur Etikettierung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N 0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Inhaltsverzeichnis

1 Zusammenfassung

Es wird ein Experimentalsystem zur automatischen prosodischen Etikettierung und Segmentierung deutscher Spontansprache vorgestellt. Das System wird eingesetzt, um eine prototypische Beschreibung prosodischer Eigenschaften (Akzentuierung und Phrasierung) einer Äußerung zu generieren. Die beschriebenen Verfahren basieren auf einer datengetriebenen statistischen Analyse der Transliteration und des zugehörigen Sprachsignals.

2 Einleitung

Aus der prosodischen Analyse gesprochener Sprache können wichtige Informationen für die automatische Spracherkennung und -interpretation und die Dialogführung gewonnen werden. Es ist aber in weiten Teilen noch nicht vollständig geklärt, inwieweit die Wahrnehmung prosodischer Phänomene durch die Sprachproduktion (z.B. Grundfrequenzänderungen), Syntax und die Semantik einer Äußerung beeinflußt werden. Ein Grund für den seltenen Einsatz prosodischer Informationen in Spracherkennungssystemen liegt in dem Mangel an prosodisch annotiertem Material begründet.

Um Untersuchungsmaterial zur Erforschung dieser Vorgänge zur Verfügung zu stellen, wird am Institut für Nachrichtentechnik der TU Braunschweig im Rahmen des Projektes Verbmobil spontansprachliches Material prosodisch segmentiert und etikettiert. Da die manuelle Segmentierung und Etikettierung eine sehr zeit- und kostenaufwendige Aufgabe ist, wird die Idee verfolgt, ein semiautomatisches System zu entwickeln.

Mit diesem System wird die Etikettier- und Segmentierarbeit unterstützt, indem für jede Äußerung automatisch eine prototypische Beschreibung der Prosodie (Akzente und Phraengrenzen) generiert wird.

Für die Erzeugung der Akzent- und Phrasengrenzenhypothesen werden statistische Methoden der Sprachmodellierung und Mustererkennung eingesetzt.

Zu diesem Zweck wird auf die Transliteration der Äußerung, die die Rechtschrift und einige zusätzliche Beschreibungen von artikulatorischen und nichtartikulatorischen Geräuschen umfaßt, und auf das Sprachsignal zurückgegriffen.

Die manuelle Segmentierung und Etikettierung beschränkt sich nach dieser Vorverarbeitung auf die *Verfizierung* der Etiketten und Segmentgrenzen. Dies senkt den Aufwand an manueller Arbeit deutlich.

Die prosodische Beschreibung beinhaltet

- Wort- bzw. Silbengrenzen als Zeitpunkte im Sprachsignal

- Zuordnungen von Akzentstufen (Nebenakzent, Hauptakzent) zu den Worteinheiten und die Lokalisierung des Akzentes im Sprachsignal durch die automatische Bestimmung des Vokalzentrums der akzenttragenden Silbe
- Bestimmung von prosodischen Phrasengrenzen

3 Systemübersicht

Das Gesamtsystem gliedert sich in drei Schritte:

1. Automatische Segmentierung des Sprachsignals in phonetische Einheiten (umfaßt somit auch Wort- und Silbengrenzen) mit einem HMM-Spracherkenner (siehe Kap 4)
2. Vorhersage von Akzenten und Phrasengrenzen aus dem Text der Äußerungen durch Bewertung mit einem stochastischen Sprachmodell (siehe Kap. 5 und Kap. 6)
3. Bestimmung der Phrasengrenzen und Akzentpositionen im Sprachsignal durch die Kombination der Ergebnisse des Verarbeitungsschrittes 1 und 2 (siehe Kap. 8.2)

Das Gesamtsystem ist ein modulares System mit klar definierten Schnittstellen. Diese Konzeption wurde gewählt, um die Austauschbarkeit der verschiedene Module zu gewährleisten. Es ist somit möglich, alternative Vorgehensweisen (z.B. bei der Akzentdetektion) schnell zu erproben, indem das entsprechende Modul ausgetauscht wird. Dieses Vorgehen verkürzt die Entwicklungszeiten neuer Algorithmen erheblich, da nicht jedesmal eine vollständig neue Integration vorgenommen werden muß. Die im Vergleich zu einem integrierten Ansatz auftretenden Geschwindigkeitseinbußen bei der Analyse sind hinnehmbar, da das gesamte System vollständig im Hintergrund (Batch-Betrieb) laufen kann.

Neben dem Sprachsignal und der Transliteration werden vom System folgende Wissensquellen genutzt:

- Regelbasierte Wissensquellen
  - Aussprachelexikon zur Erzeugung der Graphem-Phonem-Korrespondenzen
  - Grammatiklexikon zur Zuordnung der Wörter zu ihren grammatikalischen Kategorien (Verben, Adverb usw.)
  - Markierung der Silbe, die bei der Realisierung der Standardbetonung den Wortakzent trägt.

- Statistische Wissensquellen
  - Hidden Markov Modelle für die Phonemeinheiten
  - Statistische Sprachmodelle

4 Lokalisierung von Wortgrenzen im Signal

Die Wortgrenzenbestimmung kann als eine vereinfachte Aufgabe der Spracherkennung aufgefaßt werden [BRUGNARA1993], da das Erkennungsergebnis (d.h. die Wortfolge) a priori bekannt ist. Der Spracherkenner muß nur noch die Wortgrenzen im Signal detektieren. Im vorliegenden System wird zu diesem Zweck ein auf semikontinuierlichen Hidden Markov Modellen basierender Spracherkenner eingesetzt. Als Erkennungseinheiten werden kontextfreie Phonemmodelle benutzt. Aus der Phonemsegmentierung wird in einem nachgeschalteten Verarbeitungsschritt die Wortgrenzen und die Position der Silbenkerne (im Aussprachelexikon markiert) bestimmt. Für eine weitere Beschreibung wird auf [LEHNING1994],[LEHNING1994a] verwiesen.

5 Hypothesengenerierung für Phrasierung und Akzentzuordnungen

Für die Hypothesengenerierung für Phrasierungen und Akzentzuordnungen wird auf ein kategoriales n-Gramm-Sprachmodell (siehe Kap. 7) zurückgegriffen. Die Kategorien repräsentieren die grammatikalischen Wortkategorien und die Satz- bzw. Phrasengrenzen und einige spontansprachliche Phänomene, wie nichtartikulatorische Geräusche.

Zu diesem Zweck wird die Transliteration der Äußerung in einem ersten Schritt in eine Folge von grammatikalischen Kategorieren konvertiert.

Aus zwei Gründen wurde ein kategoriales Sprachmodell gewählt:

- Geringer Umfang des Trainingsmaterials: Als Trainingsmaterial standen sechshundert Äußerungen zur Verfügung. Das Lexikon umfaßt aber ca. 2000 Einträge, so daß viele Worteinheiten — im Gegensatz zur zugehörigen Kategorie — nur einmal beobachtet wurden

- Da viele prosodische Phrasengrenzen mit syntaktischen Konstituentengrenzen, die sich in ihrer Struktur — aber mit anderem Wortlaut — wiederholen, ist ein Training auf syntaktischen Einheiten günstig.

Zur Ermittlung der Phrasierung und Akzentzuordnung wird eine Graphsuche (siehe Bild 1) gestartet. Die Wahrscheinlichkeit jedes Pfades durch diesen Graphen wird mit Hilfe des kategorialen Sprachmodells berechnet (siehe Kap. 7). Das kategoriale Sprachmodell basiert auf grammatikalischen Einheiten (z.B. noun, verb), die um die jeweilige Akzentkategorie erweitert wurden (vergl. Anhang):

Beispiel: noun (unbetont), noun_NA (Nebenakzent), noun_HA (Hauptakzent)

Das Sprachmodell wurde an prosodisch etikettierten Daten trainiert. Jedes Wort des Traningsmaterials wurde dabei in die entsprechende Kategorie konvertiert und um den manuell zugewiesenen Akzent erweitert. Phrasengrenzen wurden dabei als spezielle Wortgrenzen behandelt (|).

Für eine vorgegebene Wortkette wird somit nach folgendem Schema vorgegangen:

1. Überführe die vorliegende spontansprachliche Äußerung mit den Einheiten $w_1, \ldots, w_N$ (meistens Wörter) in eine Symbolfolge $S_1, \ldots, S_N$.

2. Hinter jedem Symbol S_i kann eine Phrasengrenze PG oder eine Wortgrenze WG, die keine Phrasengrenze darstellt, folgen. Es soll jetzt die Symbolkette $Z = Z_1, \ldots, Z_L$ gefunden werden, die die höchste a priori Wahrscheinlichkeit $P(Z)$ besitzt. Die Symbolkette Z wird dabei aus der Symbolkette S unter folgenden Bedingungen abgeleitet:

   - $Z_1 = S_1$
   - $Z_L = Z_N$
   - Wenn $Z_j = S_i$, dann ist $Z_{j+1} = PG$ oder WG
   - Wenn $Z_j = PG$ oder $Z_j = WG$, dann ist $Z_{j+1} = S_{i+1}$
   - Wenn $Z_j = S_i$, dann wird Z_j ein Akzentuierungattribut zugeordnet, das eines der folgenden drei Werte annimmt:
     - <>: Leerzuweisung (Wort ist unbetont)
     - NA: Wort trägt Nebenakzent
     - PA: Wort trägt Hauptakzent

Anschaulich kann man die o.g. Bedingungen so interpretieren, daß die Symbolkette Z eine mögliche Realisierung der Symbolfolge S mit eingefügten Phrasengrenzen und Akzentuierungen darstellt. Zu einer Symbolkette S mit der Länge

N gibt es $2^{N-1}3^N$ verschiedene Symbolketten Z der Länge $2N-1$.

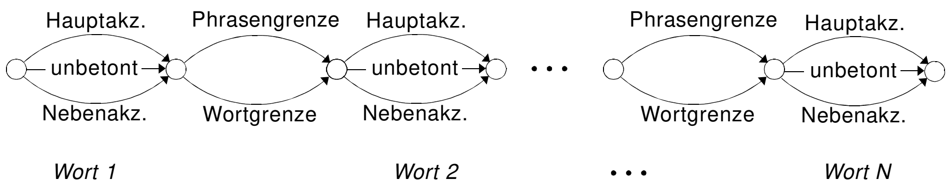

Bild 1: Darstellung der Phrasengrenzendetektion und Akzentzuordnung als Graphsuche

Es ist leicht einsichtig, daß bei längeren Symbolketten S eine vollständige Graphensuche einen enormen Suchaufwand erfordert. Deshalb wird die Suche als suboptimaler Viterbi-Beam-Search (Siehe Kap. 6) organisiert. Als Ergebnis werden die k-besten Phraseneinteilungen der Symbolkette S geliefert.

Der Einsatz eines n-Gramm-Sprachmodelles besitzt folgende Vorteile:

- Die Schätzung der Verteilungsparameter kann datengetrieben durchgeführt werden, d.h. es muß kein Regelwerk erstellt werden, in der die Abhängigkeit der Akzentuierung und der Phrasierung von der Wortstellung bzw. -folge explizit formuliert werden muß
- Aus der statistischen Schätzung kann aber ggf. ein Regelwerk abgeleitet werden, da ein m-Gramm-Modell als die Definition einer regulären Sprache intepretiert werden kann

Um die wahrscheinlichste Zuordnung zu ermitteln, ist es denkbar, alle möglichen Kombinationen von Phrasengrenzen und Akzentzuordnungen auszutesten. Dies ist aber bei längeren Wortketten praktisch nicht machbar, da bei N Wörtern $2^{N-1}3^N$ Kombinationen (s.o.) möglich sind. Aus diesem Grund wird bei der Phrasengenerierung auf Pruning-Techniken zurückgegriffen (Beam-Search), die eine suboptimale Suche darstellen.

6 Beam-Search-Technik

Beim Beam-Search handelt es sich um eine suboptimale Breitensuche in einem Graphen. Bei der Suche werden nur Verzweigungen im Graphen berücksichtigt, deren Bewertungen relativ zur besten Bewertung eine gewisse Schwelle nicht unterschreiten. Alle anderen Zweige werden verworfen (Pruning). Beim Erreichen der Blätter im Graphen werden die besten Symbolfolgen innerhalb des Graphen durch Rückverfolgung des Weges im Graphen (Backtracking) ermittelt.

6.1 Implementierung

Für die Ermittlung der Phrasengrenzen wird die Symbolfolge S von links nach rechts nach folgendem Schema abgearbeitet:

Historienliste(n = 1) mit Symbol S_1 initalisieren				
FOR $1 \leq n < N$				
	Setze $P(H_{best})$ zurück			
	FOR Für jede in der Historienliste(n) $h_1 \ldots h_K$ gespeicherte Symbolhistorie H			
		erzeuge aus der Worthistorie H sechs neue Historien		
		1. Fall: S_{n+1} trägt kein Akzent		
		$H^1 = H + S_{n+1}$ und		
		$H^2 = H + P + S_{n+1}$		
		2. Fall: S_{n+1} trägt Nebenakzent		
		$H^3 = H + S_{n+1}$ und		
		$H^4 = H + P + S_{n+1}$		
		3. Fall: S_{n+1} trägt Hauptakzent		
		$H^5 = H + S_{n+1}$ und		
		$H^6 = H + P + S_{n+1}$		
		FOR Für $i = 1, 3, 5$		
			Berechne $P(H^i) = P(H) + p(S_{n+1} \vert h_{K-1} h_K)$	
		FOR Für $i = 2, 4, 6$		
			Berechne $P(H^i) = P(H) + p(S_{n+1} \vert h_K P) + p(P \vert h_{K-1} h_K)$	
		FOR Für $i = 1, \ldots 6$		
			IF $P(H^i) > P(H_{best})$	
			THEN $P(H_{best}) = P(H^i)$	
			IF $P(H^i) > P(H_{best}) * \Theta$	
			THEN	Speichere H^i mit den a priori Wahrscheinlichkeiten in der Historienliste(n+1)
Ausgabe der wahrscheinlichsten Historie aus der Historienliste $H(N)$ mit der zugehörigen Phrasen- und Akzentzuordnung				

Bei der konventionellen Implementierung des Beam-Search wird die Beschneidung des Baumes erst nach Ermittlung der höchsten a priori Wahrscheinlichkeit durchgeführt. Dies führt aber zu deutlichen Geschwindigkeitseinbußen. Durch die adaptive Anpassung der Pruning-Schwelle werden zwar mehr Hypothesen als notwendig weiterverfolgt. Dieser erhöhte Aufwand ist aber wie in [LOWERRE1977] gezeigt vernachlässigbar.

7 Training der statistischen Sprachmodelle zur Phrasengrenzendetektion und Akzentzuordnung

Mit n-Gramm-Grammatiken bezeichnet man eine datengetriebene Methode zur Abschätzung einer Wortfolge $\mathbf{w} = \mathbf{w_1} \dots \mathbf{w_m}$ aus einem Vokabular $W = W_1, W_2, \dots, W_L$. Die Wahrscheinlichkeit eines Wortes w_k wird dabei abhängig von den Vorgängerwörtern $w_1, \dots, w_{k-1}$ geschätzt $(P(w_k|w_1, \dots w_{k-1})$. Die Wahrscheinlichkeit $P(w)$ für eine Wortfolge w ergibt sich somit zu

$$P(w) = P(w_1) \prod_{i=2}^{m} P(w_i|w_1 \dots w_{i-1})$$

Bei einer n-Gramm-Grammatik wird die Einschränkung vorgenommen, daß die Wahrscheinlichkeit $P(w_k|w_1 \dots w_{k-1})$ nur aus den maximal n vorangegangenen Wörtern geschätzt wird:

$$P(w_k|w_1 \dots w_{k-1}) \approx P(w_k|w_{k-n+1} \dots w_{k-1}$$

Für die Akzent- und Phrasengrenzenbestimmung wurde ein Trigram ($n = 3$) trainiert.

7.1 Glättungsmethoden

Das größte Problem bei dem Training des Trigram-Modells ist die ungenügende Anzahl von Trainingsdaten für die Schätzung der Modellparameter. Dies ist leicht einsehbar, wenn man sich z.B. vergegenwärtigt, daß es bei einem Wortschatz von 1000 Wörtern eine Milliarde (1000^3) mögliche Trigrame gibt. Aus diesem Grund werden die Verteilungsparameter des Trigrams geglättet.

Die Parameter werden dabei so eingestellt, daß auch im Trainingsmaterial ungesehene Wortfolgen eine gewisse Wahrscheinlichkeit zugeordnet wird. Aus diesem Grund wird anstatt einer im Sinne der Wahrscheinlichkeitsrechnung besten Abschätzung des Trainingsmaterials (Most likelihood) ein interpoliertes Modell berechnet, indem von jedem geschätzten Wahrscheinlichkeitswert ein gewisses Wahrscheinlichkeitsmaß abgezogen wird (Discounting). Die so gewonnene "Wahrscheinlichkeitsmasse" wird auf die ungesehene Ereignisse verteilt. Für das Sprachmodell zur Akzent- und Phrasengrenzengenerierung wurde als Glättungsverfahren die nichtlineare Interpolation mit absolutem Abzugswert (Discounting Value) gewählt.

Bei diesem Verfahren wird von der absoluten Beobachtungshäufigkeit $N(k)$ (ermittelt durch Zählung) eines Trigrams k ein Betrag D (Discounting Value) abgezogen [NEY1994]. Häufig wird D zu 1.0 gesetzt, d.h. anschaulich, daß es für die Schätzung der Auftrittswahrscheinlichkeit gleichgültig ist, ob ein Trigram einmal oder nie beobachtet wurde.

Die Berechnung der Trigram-Wahrscheinlichkeiten wird dann in folgenden Schritten ausgeführt (unter der Annahme $D \leq 1$):

1. Berechne $p(w_M|w_{M-2}w_{M-1})$

$$\frac{max(0, N_{w_M,w_{M-1},w_{M-2}} - D)}{N_{w_{M-1},w_{M-2}}} + D\frac{L - W_0^3(w_{M-1}, w_{M-2})}{N_{w_{M-1},w_{M-2}}}p(w_M|w_{M-1})$$

2. Berechne $p(w_M|w_{M-1})$

$$\frac{max(0, N_{w_M,w_{M-1}} - D)}{N_{w_{M-1}}} + D\frac{L - W_0^2(w_{M-1})}{N_{w_{M-1}}}p(w_M)$$

3. Berechne $p(w_M)$

$$\frac{max(N_{w_M} - D, 0)}{N} + D\frac{L - W_0}{N}p(Zerogram)$$

4. Berechne $p(Zerogram) = \frac{1}{L}$ $\quad L$: Vokabulargröße

Erläuterungen:

L	:	Vokabulargröße
W_0	:	Anzahl der Wörter, die nie beobachtet wurden (max. M)
$W_0^2(w_M)$	:	Anzahl der Wortpaare w_{M-1}, w_M, die mit w_M enden und nie beobachtet wurden (max. M)
$W_0^3(w_M, w_{M-1})$	:	Anzahl der Wort-Dreiergruppen w_{M-2}, w_{M-1}, w_M, die mit w_{M-1}, w_M, enden und nie beobachtet wurden
$N_{w_M,w_{M-1},w_{M-2}}$	:	Anzahl der gezählten Wort-Dreiergruppen, die die Wortfolge w_{M-2}, w_{M-1}, w_M enthalten
$N_{w_M,w_{M-1}}$	:	Anzahl der gezählten Wortpaare, die die Wortfolge w_{M-1}, w_M enthalten
N_{w_M}	:	Anzahl der Beobachtungen des Wortes w_M

8 Bestimmung der Signalpositionen für Akzente und Phrasengrenzen

8.1 Phrasengrenzen

Die Positionen der Phrasengrenzen im Signal werden auf das Wortende des jeweils vorangehenden Wortes gesetzt.

8.2 Bestimmung der Signalposition für Akzente

Die Akzentposition eines Wortes im Signal wird durch ein zweistufiges Vorgehen gefunden. Im ersten Schritt wird ein Viterbi-Alignment der phonetischen Wortbeschreibung auf den entsprechenden Signalabschnitt durchgeführt (siehe Kap. 4). Im zweiten Schritt wird aus dem Aussprachelexikon das Phonem ermittelt, dem die lexikalische Wortbetonung zugeordnet ist. Die Akzentposition wird dann auf das Zentrum dieses Phonems im Zeitsignal gesetzt. Die Position des Zentrums wird aus den Phonemgrenzen dieses Phonems berechnet.

9 Auswertung

Die vorhergesagten Akzente und Phrasengrenzen wurden mit manuell etikettierten Sprachmaterial verglichen. Das Vergleichsmaterial umfaßte 397 Wörter (21 Äußerungen). Die Auswertung ergab dabei die folgenden Verwechselungsmatrizen.

Akzente		
Klasse	klassifiziert als	
	nicht akz.	akzentuiert
nicht akz.	277	46
akzentuiert	18	56

Klassifikationsrate: 83.88 %

Phrasengrenzen (PG)		
Klasse	klassifiziert als	
	keine PG	PG
keine Phrasengrenzen	297	18
Phrasengrenzen	20	62

Klassifikationsrate: 90.43 %

Das Auswertungsergebnis zeigt, daß Akzente und Phrasengrenzen mit Hilfe des statistischen Sprachmodells in den meisten Fällen korrekt vorhergesagt werden können. Die Lokalisierung der Akzentpositionen im Signal wird mit einer durchschnittlichen Abweichung von ca. 10 ms gefunden, wenn zuvor die automatisch gesetzten Wortgrenzen manuell verifiziert wurden.

A Beispiel zur Phrasierung und Akzentzuordnung

```
Transliteration:
 dann w"urde ich sagen da"s wir noch ein' Termin ausmachen  <breathing>

Konvertierung in grammatikalische Kategorien:
  adv verb pro verb konj pro adv det noun verb <breathing>

Ergebnis der Vorhersage:
  adv verb pro verb_PA |  konj pro adv det noun_SA verb | <breathing>

Konvertierung der grammatikalischen Einheiten auf die Textebene:
  dann w"urde ich sagen_<PA> |  da"s wir noch ein' Termin_<PA> ausmachen |
<breathing>
```

Literatur

[BRUGNARA1993] F. Brugnara et al.: *Automatic segmentation and labeling of speech based on Hidden Markov Models*, Speech Communication 12 (1993), 357 - 370

[DEROUAULT1986] A.M. Derouault, B. Merialdo: *Natural Language Modeling for Phoneme-to-Text Transcription*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 8, pp. 742-749, Nov. 1986

[KATZ1987] S.M. Katz: *Estimation of Probalities form Sparse Data for the Language Model Component of a Speech Recognizer*, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 35, pp. 400-401, March 1987

[KOHLER1994] K. Kohler et al.: Handbuch zur Datenaufnahme und Transliteration in TP 14 von Verbmobil — 3.0, Verbmobil Technische Dokument Nr. 11, September 1994

[LOWERRE1977] B. Lowerre: *The Harpy Speech Recognition System*, PhD thesis, Carnegie Mellon University, Computer Science Dept., 1977

[LEHNING1994] M. Lehning: *Automatische Wortsegmentierung mit Semi Continuous Hidden Markov Modellen (SCHMM)*, VERBMOBIL — Technisches Dokument Nr. 4

[LEHNING1994a] M. Lehning: *Automatische Wortsegmentierung mit semikontinuierlichen Hidden Markov Modellen*, Fortschritte der Akustik, DAGA 94, Dresden, Teil C, 1257 - 1260

[NEY1994] H. Ney et al.: *On Structuring Probalistic Dependences in Stochastic Language Modelling*, Computer, Speech and Language, Vol. 8, 1994, 1 - 38

[REYELT1994] M. Reyelt, A. Batliner: *Ein Inventar prosodischer Etiketter für Verbmobil*, Verbmobil Memo Nr. 33, 1994