Überprüfung von sprachwissenschaftlichen Hypothesen mittels Sprachsynthese

Eric Keller
LAIP-IMM, Faculté des Lettres
Université de Lausanne, 1015 Lausanne, Suisse
eric.keller@imm.unil.ch

Webseite für diesen Vortrag (Beispiele) / Page Internet pour cette présentation (exemples):

www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_dt.htm

www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_fr.htm

Übersicht

1. ***Einführung:*** Kurz unsere Sprachsynthese vorstellen
2. ***Theoretische Frage:*** Benötigen wir für verschiedene Dialekte, Soziolekte und Sprachstile mehrere Prosodietheorien, oder kommen wir mit einer prototypischen Theorie plus Abweichungen aus?
3. ***Ein kleines konkretes Beispiel:*** Die Bestimmung der prosodischen Schnittstellen in der Spontansprache, verglichen mit einer neutralen gelesenen Sprache.

Organisatorisches:

1. Das konkrete Beispiel ist die gleiche Demonstration, die ich gestern auf französisch vorgestellt habe.
2. Aus Zeitgründen wurde dieser Vortrag gekürzt. Der volle Vortrag und die Tonbeispiele sind hier erhältlich:
   - www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_dt.htm
   - www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_fr.htm

Ansatz

- Moderne Sprachsynthesen erlauben es, sprachwissenschaftliche Hypothesen hörbar zu testen.
- Auf diese Art kann man z.B….
  - offensichtliche Fehler einer Hypothese identifizieren,
  - die Qualität und Vervollständigung einer Sprachbeschreibung kontrollieren,
  - Interaktionen zwischen Arbeitsabläufen untersuchen,
  - die Relevanz von Arbeitsprozessen in unterschiedlichen linguistischen und kommunikativen Kontexten überprüfen,
  - u.s.w.

Überblick der Probleme

- **Aber der Einsatz von sprachsynthetischen Methoden ist nicht ganz einfach.**
- **Technische Probleme:**
  - **Unzugänglichkeit:**
    - **Manche Parameter sind nicht zugänglich oder sind von dritten Parametern abhängig.**
  - **Unzulänglichkeiten:**
    - **Bei genauem und längerem Hinhören trügt noch keine Synthese das menschliche Ohr.**
- **Probleme bei der Interpretation der Resultate:**
  - **Die Kontamination** der Ausgabe durch zeitlich folgende Prozesse:
    - **Untersuchungsparameter wie Grundfrequenz oder Timing werden oft durch eine schlechte Syntheseausgabe verschleiert, verdeckt oder verändert.**
  - **Perzeptuelle Interpretation:**
    - **Es besteht oft Unklarheit, ob die eingegebenen Parameter unzulässig oder leicht abnormal sind, oder gar in den Rahmen der natürlichen Varianz fallen.**

Um diese Probleme einigermassen in den Griff zu kriegen, haben wir unsere eigene Synthese gebaut →

Das LAIP Syntheseprojekt (1992-heute)

Datum: Fertigstellung der Technologie

- **Themenbereiche:**
  - *Prosodie* (1992- , E. Keller, B. Zellner Keller, SNF, BBW) — **1997- Aufbau/Prosodie**
  - *Sprachstiluntersuchungen* (1998- , B. Zellner Keller, BBW) — **(in Entwicklung)**
  - *Dialektologie* (1998- , B. Siebenhaar, SNF) — **(in Entwicklung)**
  - *Signalgeneration mit Spektralmethoden* (2002- , E. Keller, BBW) — **2002 - HNM / F0 - Dauer**
- **Sprachsynthesesysteme**
  - LAIPTTS-F (Französisch, 1994-1997, E. Keller, B. Zellner Keller, Mitarbeiter, SNF, BBW) — **1997**
  - LAIPTTS-D (Hochdeutsch, 1998-2002, B. Siebenhaar, M. Forst, BBW) — **2002**
  - LAIPTTS-SwD (Schweizerdeutsch, 2001- , B. Siebenhaar und Mitarbeiter-innen, SNF) — **(in Entwicklung)**
  - LAIPTTS-L (Latein, 2001- , O. Bianchi) — **2001**
  - LAIPTTS-E (UK-Englisch, 2002- , E. Keller) — **(in Entwicklung)**

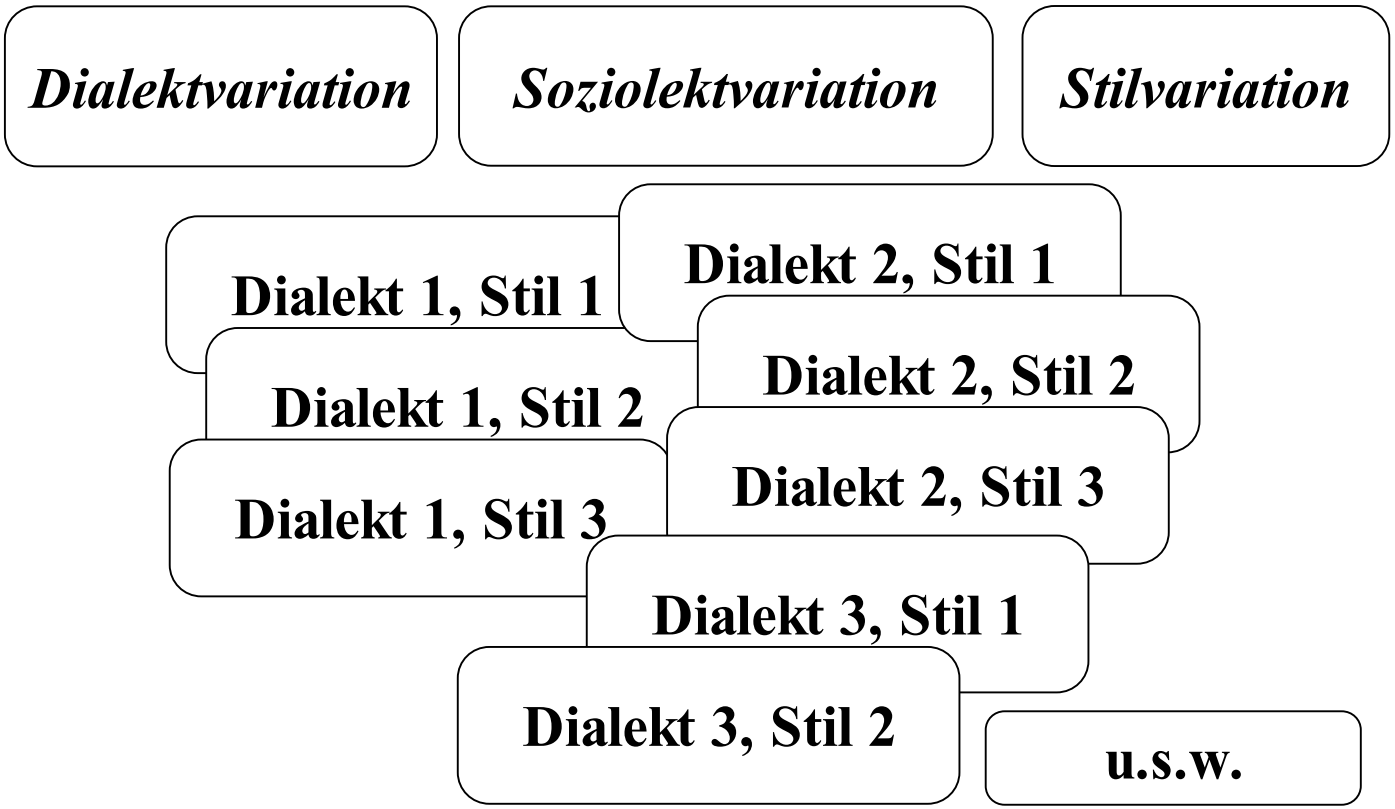
Ganz kurz: LAIPTTS Prosodie

- **Grundlagen:** Wir benutzen psycholinguistisch und statistisch motivierte Modelle für Phrasierung, Timing und Intonation.
- Die **Eingabe** ist eine phonetische Lautkette, mit Wort- und Silbenmarkierungen plus Zeichen für den grammatischen oder lexikalischen Status von Wörtern.
- Diese Lautkette wird in **prosodische Haupt- und Nebenphrasen** unterteilt.
- Auf Grund von etwa 15 Parametern wird die **Segmentdauer** bestimmt (z.B. phonetische Identität des zu bestimmenden, des vorhergehenden und darauffolgenden Segments, Anzahl der Silben, Position der Silbe im Wort, u.s.w.).
- Die **Grundfrequenz** wird mit einem additiven Fujisaki-Modell berechnet. Kurven werden in diesem Modell aus einer Kombination von Akzentimpulsen und Phrasenimpulsen berechnet.
- Wir haben (wie verschiedene andere Forscher) gezeigt, dass **stochastische Methoden** (Statistik, neuronale Netze) zur Zeit noch genaure Berechnungen von Zeit und Grundfrequenz machen als phonologische Modelle (Keller & Zellner, 1995, Zellner, 1998).

Was uns zur Zeit beschäftigt: Wieviel Prosodiemodelle benötigen wir?

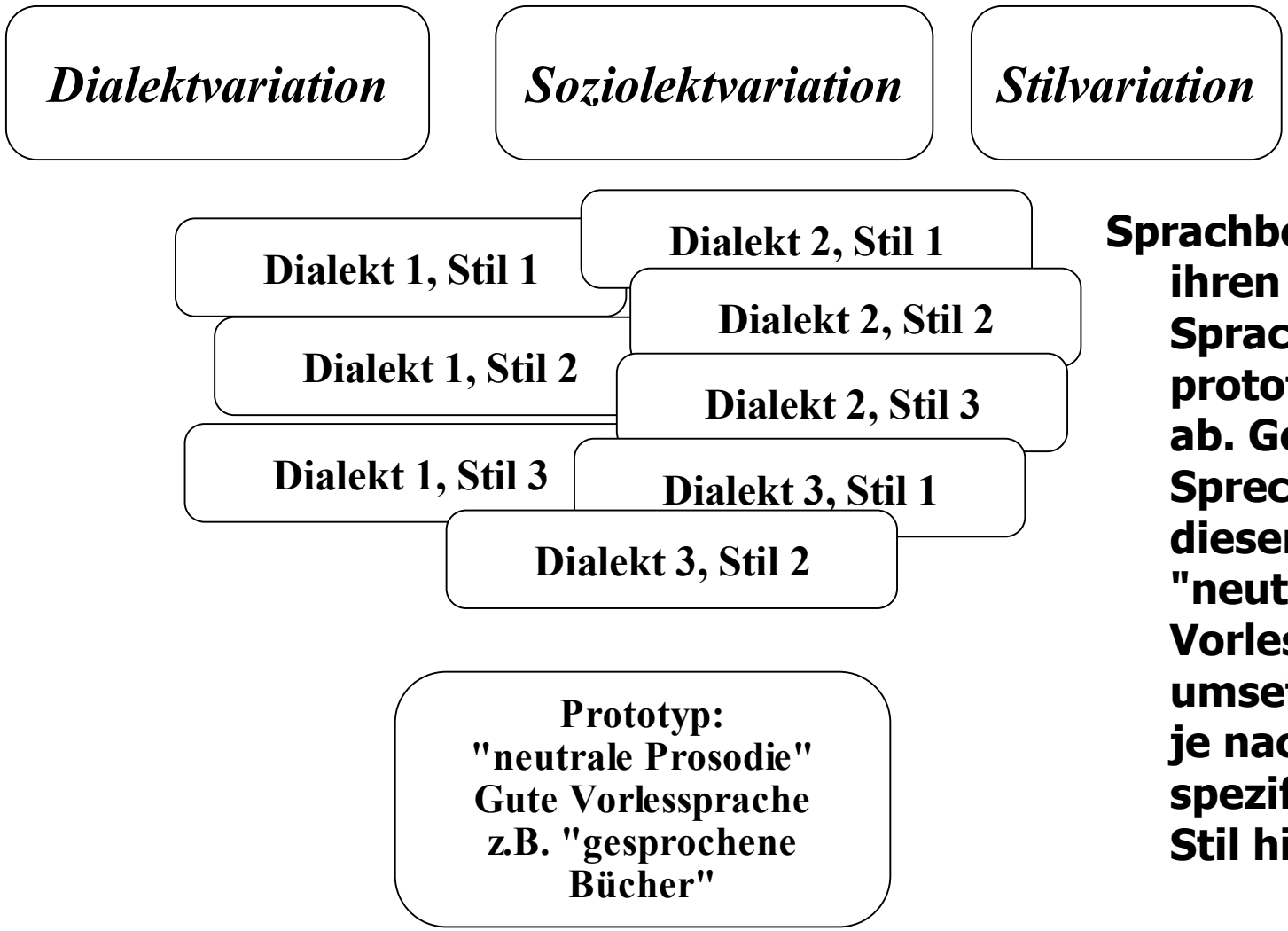
- Welches sind die Unterschiede bei der Zeit- und Grundfrequenzbestimmung in *verschiedenen* Sprachen, Dialekten und Sprachstilen?
- Können Zeit und Grundfrequenz für verschiedene Dialekte und verschiedene Sprachstile mit ähnlich strukturierten Systemen berechnet werden? Oder benötigen wir für jede Sprache, jeden Dialekt, Soziolekt und Stil eine neue Struktur?
- Drei Modelle sind denkbar:
  1. Ein distributives Modell
  2. Ein prototypisches Modell
  3. Ein performanzgetriebenes prototypisches Modell

1. Ein distributives Modell



Die Variation versteht sich als teilweise Überlagerung von verschiedenen Prosodiemodellen. Je nach Umständen wird ein bestimmter Dia-Soziolekt und Sprechstil gewählt. (Der Einfachheit halber illustrieren wir hier nur Dialektvariationen.)

2. Ein prototypisches Modell



Sprachbenutzer leiten aus ihren Spracherlebnissen eine prototypische Prosodie ab. Geübte Leser-Sprecher können diesen Prototyp in eine "neutrale" Vorlessprache umsetzen, und können je nach Bedarf in einen spezifischen Lekt und Stil hinüberwechseln.

Performanzbegrenzungen als Eckpfeiler der Prosodie

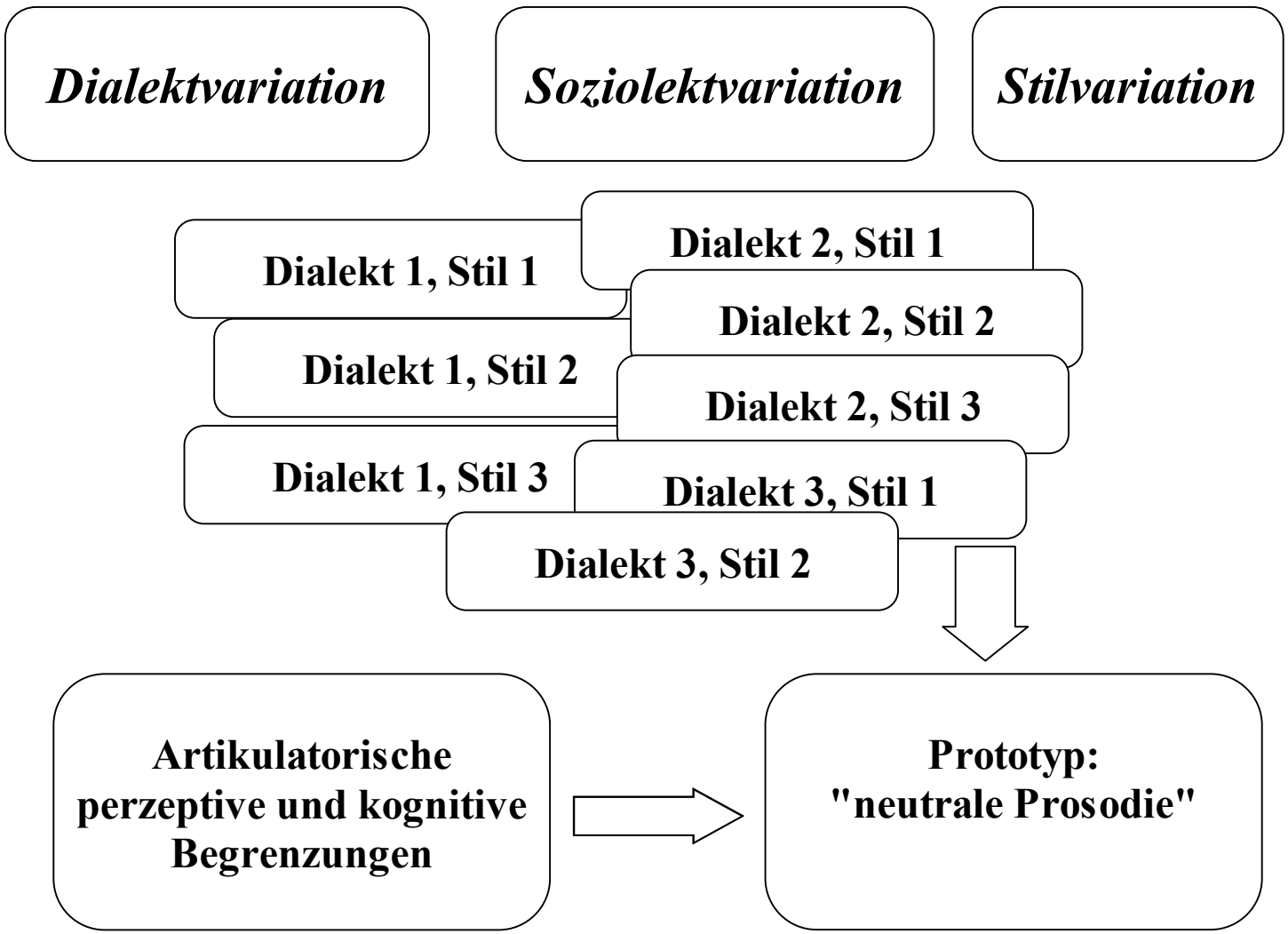
- Die Prototypenbildung wird wahrscheinlich durch universelle artikulatorische, perzeptive und kognitive Begrenzungen unterstützt.

  Beispiele:

  - Veränderungen der ***Grundfrequenz*** sind in Bezug auf Höhe, Tiefe und Geschwindigkeit begrenzt. Dies begrenzt die Variation der F0.
  - Die distinktive Natur der phonetischen Kodierungn auferlegt jedem Sprachlaut eine gewisse ***Mindestdauer,*** und die ***Maximaldauer*** wird durch den Kommuniationsdruck begrenzt. Dies begrenzt die Variation der Segment- und Silbendauer.
  - Die ***Dauer von Segmenten*** wird hauptsächlich durch die phonetische Identität des Segmentes und durch diejenige der kontextuellen Segmente bestimmt. Dies begrenzt die mögliche Struktur des Modells (Arbeiten von Campbell, Traber, Keller, Zellner, u.s.w.)
  - Äusserungen, länger als 10-14 Silben ohne Zäsur (Pausen, abrupte Intonations- und Dauerveränderungen), sind schwer zu verstehen. Dies begrenzt die Länge von Phrasen (Konzept der ***"Performanzstrukturen"***, siehe Arbeiten von F. Grosjean, UniNE, und die Überarbeitung dieser Konzepte von B. Zellner, UNIL).

- Solche Begrenzungen sind schlussendlich physiologischer Natur und können deshalb sprachunabhängig in jedes Prosodiemodell miteinbezogen werden. Untersuchungen in unserem Labor haben gezeigt, dass auf diese Art im Französischen und Deutschen, wie auch in der raschen und langsameren Sprache, relativ invariable prosodische Äusserungsparadigmen eingesetzt werden können.

3. Ein performanzgetriebenes prototypisches Modell



Der Prototyp wird unter Beachtung der Performanz-begrenzungen aus der Variation abgeleitet.

Eine Probe aufs Exempel: prosodische Schnittstellen in der Spontansprache

- Wie würden wir vorgehen, um die prosodischen Schnittstellen ("Zäsuren") in der Spontansprache…
  1. Zu idenfizieren?
  2. Auf Grund einer Transkription zu simulieren?
- *In anderen Worten:* Können die Prinzipien der "neutralen Prosodie" auf die Spontansprache übertragen werden?

Analyse: Fangen wir vorne an →

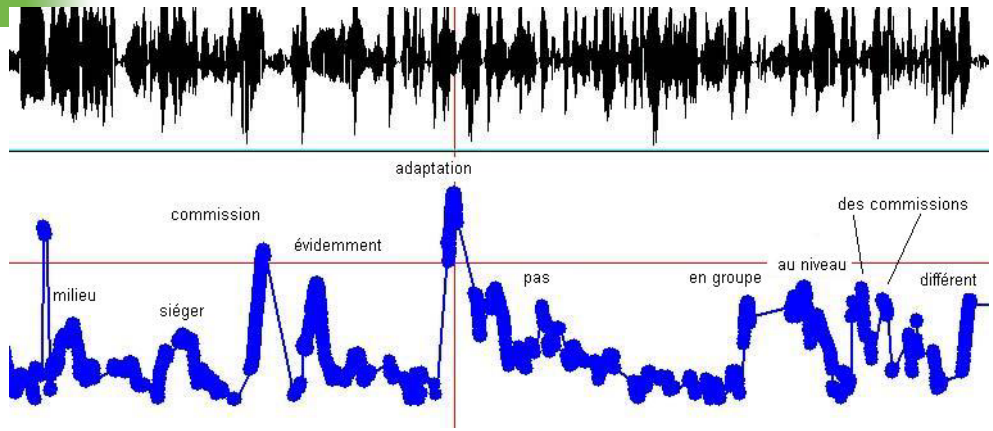
Wo sind die Schnittstellen in dieser Spontansprache?

- **Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a une certain temps d'adaptation si au plenum ça ne joue pas véritablement son rôle parce que le décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent | on a pu assister parfois où lorsque le nouveau député arrive en cours de législature pendant plusieurs séances il n'a pas le même poids parce qu'il ne connaît pas le dossier | ça se comprend effectivement et souvent le député n'arrive pas à s'affirmer assez rapidement contre le gouvernement et en cours de législature vous doutez bien que ça sert le gouvernement.** (| = Schnittstellen, die durch die Begrenzungen unserer Synthese bedingt wurden. In blau, der Teil, den wir analysieren werden).

- **Der Abgeordnete der in der Mitte der Legislaturperiode ankommt und der einer Kommission angehört braucht logischerweise eine gewisse Adapationszeit wenn dies im Plenum keine grosse Rolle spielt weil die Entscheidungen oft gemeinsam gefällt werden ist das in den Kommissionen doch ein bisschen anders es hat Fälle gegeben wo der Abgeordnete der in der laufenden Legislaturperiode ankommt während mehreren Sitzungen nicht das gleiche Gewicht hat weil er die Akten nicht kennt das versteht sich klar und oft kann sich der Abgeordnete nicht gegen die Regierung behaupten und Sie verstehen dass dies der Regierung in der Legislaturperiode zugute kommt.** (Der ehemalige jurassische Abgeordnete Claude Laville [CSI], RSR 20.10.02)

Perzeptiver Eindruck: die Intonation liefert die Indizien

- ***Perzeptiv markiert:*** **Le député qui arrive | au milieu de législature | et qui doit siéger dans une commission | évidemment | il a un certain temps d'adaptation | si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe | au niveau des commissions | alors c'est un tout petit peu différent. |**



- **Original**
- **"Hum", generiert auf Grund einer Grundfrequenzextraktion (Praat)**

- ***Durch F-Null markiert:*** **Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a un certain temps d'adaptation si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent.**

- **Übereinstimmungen (commission, adaptation, groupe, différent): 4**
- **Nichtidentifikationen von perzipierten Markierungen (arrive, législature, évidemment, commissions): 4**
- **Falsche Identifikationen von unmarkierten Silben (milieu, siéger, évidemment, pas, niveau, des, commissions): 7**

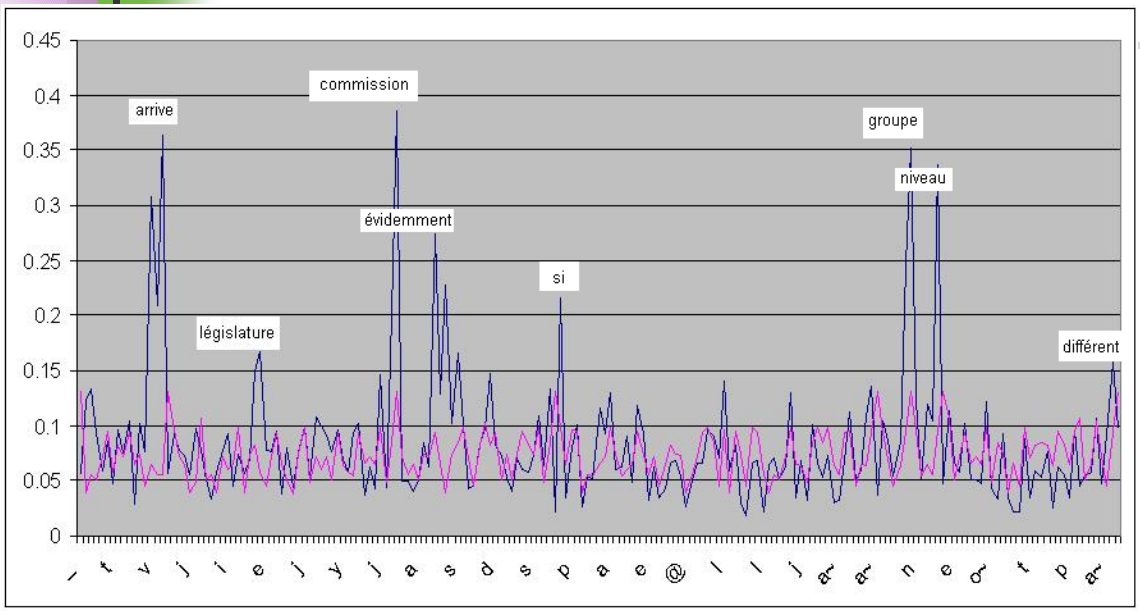
Erste Reihe von Beobachtungen

- Nicht- und falsche Identifikationen sind häufig.
- F0-Gipfel markieren oft Silben in der Mitte oder am Anfang von Wörtern.

→ Der akustische Parameter für die Melodie (die Grundfrequenz oder f-Null), alleine genommen, kann also nicht als zuverlässiges Indiz für die perzipierten Schnittstellen gelten.

Liefert die Dauer die benötigten Indizien?

- ***Perzeptiv markiert:*** **Le député qui arrive | au milieu de législature | et qui doit siéger dans une commission | évidemment | il a un certain temps d'adaptation | si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe | au niveau des commissions | alors c'est un tout petit peu différent. |**



- **In blau: gemessene Segmentdauer von M. Laville**
- **In rot: Durchschnitts-segmentdauer in einem Korpus von 12'000 Segmenten (ebenfalls ein jurassischer Sprecher, Prof. A. Wyss)**

- ***Durch Segmentdauer markiert:*** **Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a un certain temps d'adaptation si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent.**

- **Übereinstimmungen (arrive, législature, commission, évidemment, groupe, différent): 6**
  - **Nichtidentifikationen von perzipierten Markierungen (adaptation, commissions): 2**
    - **Falsche Identifikationen (si, niveau): 2**

Zweite Reihe von Beobachtungen

- Starke Verlangsamungen berühren typischerweise Schlusssilben, direkt vor den perzipierten Schnittstellen.
- Nicht- und Falschidentifikationen sind seltener.
- Übereinstimmungen sind häufig.
- B. Zellner (1998, 2002) hat gezeigt, dass eine Silbendauer von mehr als einer Standardabweichung in einem "neutralen" Text verlässlich Hauptabschnitte identifiziert.

→ Eine auf Temporalstrukturen basierte Markierung scheint also interessant.

- Die Falschidentifikationen:
  - "si" könnte hinzugewonnen werden, wenn man Schnittstellen *vor* starken Verlangsamung ebenfalls anerkennen würde.
  - "niveau" könnte als Zögern interpretiert werden (wobei eingeräumt wird, dass dies die Frage der Unterscheidung zwischen Zögern und Verlangsamung aufwirft).
- Eine komplexe Kombination zwischen zeitlichen und melodischen Indizien wäre ebenfalls denkbar.

Simulation basierend auf einer temporalen Schnittstellendefinition

- ***Durch Segmentdauer markiert:*** **Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a un certain temps d'adaptation si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent.**

- **Sprachsynthese für Französisch:** LAIPTTS-F.

  - **Mit Mbrolaausgabe:** 

Und wenn wir vom Text ausgingen (TTS-Modus)?

- Bis jetzt haben wir Kopie-Synthese betrieben, d.h., wir haben einfach die beobachteten Indizien wieder in den Text hineingegeben.
- Könnte man die Schnittstellen direkt von der Transkription ableiten?
- Das wäre die normale Arbeitsweise in TTS-Systemen.

Was wissen wir über Text-basierte Schnittstellen?

- Wir wissen, dass neutral gelesene Texte durch eine starke Regelmässigkeit gekennzeichnet werden (Zellner, 1998):
  - *An Punkten und Kommas*, Hauptschnittstellen.
  - *Wenn ein Abschnitt 12-14 Silben übersteigt*, zusätzliche Schnittstellen ("performance groups")
  - *Normalerweise, Erhaltung der Gruppe G(G….)L(L…)*, wo G = grammatisches Wort, und L = lexikalisches Wort

    $\rightarrow$ Schnittstellen nur an der L|G Transition
- Diese Regelmässigkeiten erlauben uns, gelesene Sätze mit einem Minimum von syntaktischem Wissen zu generieren.
- Respektiert die Spontansprache die gleichen Regeln? Schauen wir unseren Spontantext an.

Wo befinden sich die Schnittstellen, die durch starke Verlangsamungen gekennzeichnet werden?

- **Le député qui arrive**
- **au milieu de législature**
- **et qui doit siéger dans une commission**
- **évidemment**
- **il a un certain temps d'adaptation**
- **si au plenum ça n'joue pas véritablement son rôle ^ parce que les décisions sont souvent prises en groupe**
- **au niveau**
- **des commissions alors c'est un tout petit peu différent.**

- **GLGL 8 Silben**
- **GLGL 8 Silben**
- **GGGLGGL 10 Silben**

- **L 4 Silben**
- **GGGLLL 10 Silben**

- **GGLGGLGLGLGGGLGLLGL**
  26 Silben(14 + 12)
- **GL 3 Silben**
- **GLGLGGGGLGL**
  14 Silben

Dritte Reihe von Beobachtungen

(Wir nehmen hier an, dass eine starke Verlangsamung empirisch eine Schnittstelle markiert.)

- Eine Schnittstelle zwischen zwei lexikalischen Wörtern wird als Satzschnittstelle (hier eine Apposition) interpretiert (L-L Transition = äquivalent zu einem Schlusspunkt).
- Eine Schnittstelle zwischen einem lexikalischen und einem grammatischen Wort wird als Phrasierungschnittstelle interpretiert. (L-G Transition = äquivalent zu einem Komma).
- Wir beobachten keine einzige Schnittstelle an einer G-L Transition.
- L-G Transitionen sind also besonders "wahlfähig" für eine Schnittstelle.
- Wir haben einen einzigen Abschnitt mit mehr als 14 Silben gefunden. In diesem Falle wäre eine Unterteilung in zwei kürzere Abschnitte (von 14 und 12 Silben) auch denkbar.

Spekulation und Simulation

- *Frage:* Was würde sich nach dem folgenden absurd vereinfachten Zäsurmodell ergeben…
  - Ein Abschnitt alle ±14 Silben?
  - Jeweils an der nächstliegenden L-G Schnittstelle?

Le député qui arrive au milieu de législature | et qui doit siéger dans une commission évidemment | il a un certain temps d'adaptation si au plenum | ça n'joue pas véritablement son rôle parce que les décisions | sont souvent prises en groupe au niveau des commissions | alors c'est un tout petit peu différent. |

- Nicht schlecht, ausser dem Fehlen der Schnittstelle nach "groupe", das den Sinn verändert.

Für Skeptiker:
Simulation des restlichen Textes nach den gleichen Prinzipien

- On a pu assister parfois où lorsque | le nouveau député arrive en cours de législature | pendant plusieurs séances il n'a pas le même poids | parce qu'il ne connaît pas le dossier.
- Ça se comprend effectivement et souvent | le député n'arrive pas à s'affirmer assez rapidement | contre le gouvernement et en cours de législature | vous doutez bien que ça sert le gouvernement.

Folgerungen

- Wir behaupten nicht, das Ei des Kolumbus gefunden zu haben! Die Semantik, Pragmatik, und Syntax werden an gewissen Orten passendere Zäsuren auferlegen, als diejenigen, die unser stark vereinfachtes Ausgangsmodell vorgibt.
- Aber wir haben die Glaubhaftigkeit dreier Konzepte dargelegt:
  - Die Dominanz der Temporalstrukturen für die empirische Markierung von Schnittstellen
  - Die Empfindlichkeit der L-G Transition
  - Die approximative Begrenzung von ±14 Silben der Performanzstrukturen
- Wir denken, dass dies die wichtigsten Linien einer phonologischen Struktur darstellt, auf die die Wahl der Zäsur aufgebaut wird.

→ Analogie

Analogie

- Die Struktur der "phonologischen Zäsurempfindlichkeiten" ist wie die Form eines Glases, in welches man den "Sinn" giesst. Das Glas auferlegt dem Sinn seine Form.
- In anderen Worten, der Sprecher wird versuchen, so gut wie möglich die prototypisierten phonologischen Gegebenheiten der Sprache auszunutzen, um seinen Sinn zu übertragen.
- Und diese prototypischen Gegebenheiten sind die gleichen in der gelesenen wie auch in der spontanen Sprache.

Danke für Ihre Aufmerksamkeit!

Referenzen

- Internetseite für diesen Vortrag (Vollversion + Beispiele):

  www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_dt.htm

  www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_fr.htm

- Bibliografische Referenzen:

  - **Keller, E., & Zellner, B. (1998).** Motivations for the prosodic predictive chain. Proceedings of ESCA Symposium on Speech Synthesis. Paper 76, pp. 137-141. Jenolan Caves, Australia. Erhältlich: www.unil.ch/imm/docs/LAIP/Kellerdoc.html.
  - **Siebenhaar-Rölli, B., Zellner Keller, B., & Keller, E. (2001).** Phonetic and Timing Considerations in a Swiss High German TTS System. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (eds.). Improvements in Speech Synthesis (pp. 165-175). Wiley & Sons.
  - **Zellner Keller, B. (2002).** Revisiting the Status of Speech Rhythm. in Bernard Bel & Isabelle Marlien (eds.), 2002. Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002.(pp. 727-730). Aix-en-Provence: Laboratoire Parole et Langage. ISBN 2-9518233-0-4.
  - **Zellner Keller, B. (2002).** La simulation du rythme de parole. Travaux de l'Institut de Phonétique de Strasbourg. TIPS 31 (pp. 139-165). ISDN 0750-1315.
  - **Zellner, B. (1998).** Caractérisation et prédiction du débit de parole en français. Une étude de cas. Thèse de Doctorat. Faculté des Lettres, Université de Lausanne. Erhältlich: www.unil.ch/imm/docs/LAIP/ZellnerKellerdoc.htm.