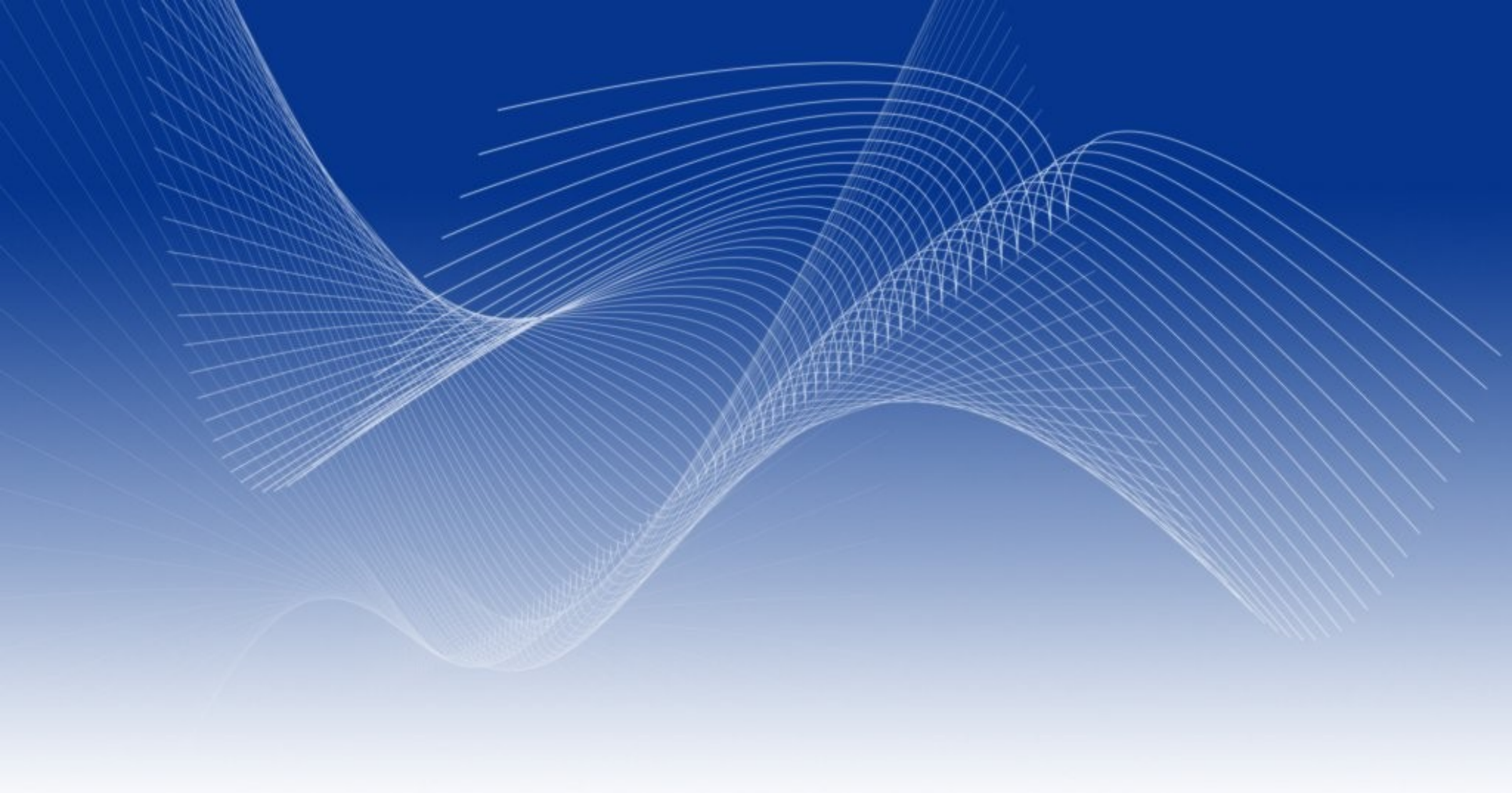


# Automatische und universelle Analyse und Segmentierung von Strings natürlicher Sprachen auf der Ausdrucksebene

2006-10-25

**Markus Stengel**

# Inhalt

- Einführung
- Meta-Rating
- Versuchsaufbau
- Exzerpieren von Korpora
- Detektion syntaktischer Trenner
- Textsegmentierung
- Fazit
- Ausblick

# Einführung

- Ziel: universelle Analyse der Sprache über Schrift
- wichtig: Bedeutung wird nicht beachtet
- Szenario:
  - Buch in unbekannter Sprache gegeben
  - Vokabular erstellen
- stark vereinfachend: syntaktische Trenner (Leerzeichen, Kommata, Punkte, etc.)
- einfach für:
  - z.B. Deutsch, Englisch, Hebräisch
  - “Heute findet wo ein Fest statt?”
- schwierig:
  - z.B. Japanisch, Chinesisch
  - “今日はどこに祭があるか。“

# Einführung

- Mechanismen und Eigenschaften einer Sprache können die Analyse der Schrift erschweren:
  - fehlende syntaktische Trenner
  - Umfang Vokabular (Englisch, Japanisch)
  - Affixation mit Einfügen von Buchstaben (“sin**n**er”)
  - Uneinheitliche Konjugationen (“putzen”=>”putzte”, “laufen”=>”lief”)
  - Groß-/Kleinschreibung (“Ausgehen” <-> “ausgehen”)
  - Komposita (“Dampf”+”Schiff”=>”Dampf**s**chiff”)
  - Vokalmodifikationen (“Haus” => “H**ä**user”)
  - orthographische Variation (“ 水圧“ , “ すいあつ“ )
  - Mehrfachbedeutungen (“ の“ ,“ に“ ,“ は“ )
  - Abkürzungen (“ 外国銀行” =>” 外行” )

# Versuchsaufbau

- Korpora:
  - Bibelkorpora (**fett**) von besonderer Bedeutung:
    - Vergleichbarkeit
    - Vokabular deckt Alltagssprache gut ab
  - drei englische: 143.541, 613.080, **4.024.663** Zeichen
  - fünf deutsche: 702.172, 1.363.272, **4.024.851**, 4.890.795, 4.940.711 Zeichen
  - zwei hebräische: **1.624.993**, **2.194.049** Zeichen
  - fünf japanische: 1.244.156, 544.771, **1.788.927**, 141.203, 10.263.491 Zeichen
- Aufbereitung der Korpora:
  - Entfernung von Zeilenumbrüchen
  - Ersetzung mehrfacher Leerzeichen durch ein einzelnes

# Bewertung und Meta-Rating

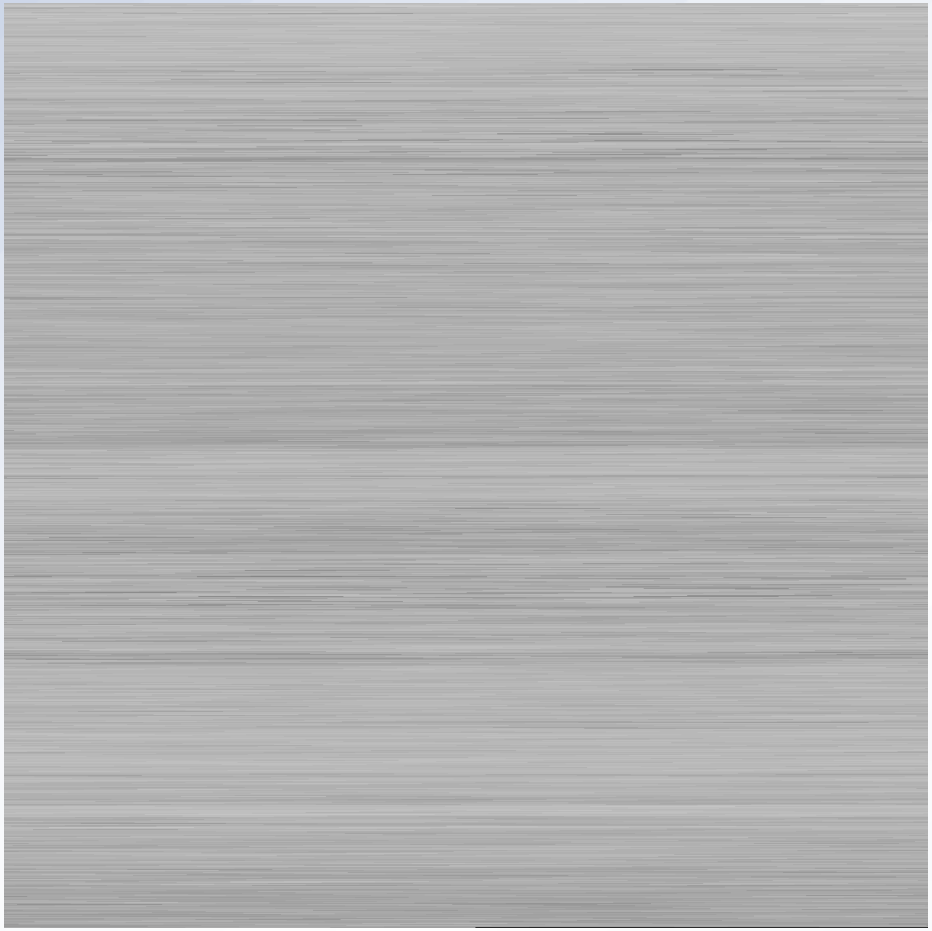
- Keine absoluten Werte als Thresholds, nur relative; Ausnahmen: 0, 1, 50%
- Entwicklung eines **Meta-Rating**:
  - ermöglicht Vergleichbarkeit unterschiedlicher Methoden
  - arbeitet nur mit berechneten Resultaten, keine vorgegebenen Werte
  - anwendbar auf sich selbst
  - Idee: Normalisierung auf Intervall [0,1] über Differenz Maximum minus Minimum, 1,0 ist bestes Ergebnis
- Beispiel:

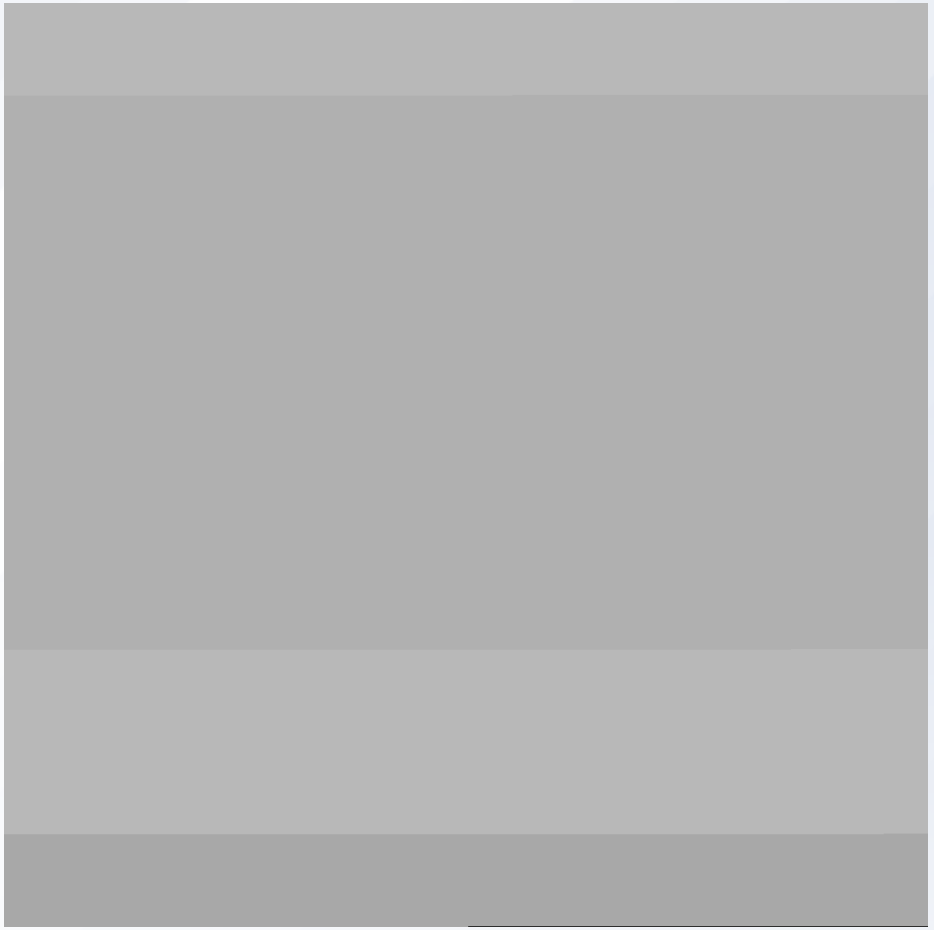
| Zeichen | Meth. 1 | Meth. 2 | Meth. 3 |
|---|---|---|---|
| „e“ | 0.900 (1) 1.000 | 0.051 (1) 1.000 | 100 (2) 0.000 |
| „ “ | 0.897 (2) 0.996 | 0.045 (2) 0.714 | 200 (1) 1.000 |
| „a“ | 0.100 (3) 0.000 | 0.030 (3) 0.000 | 100 (2) 0.000 |

# Exzerpieren von Korpora

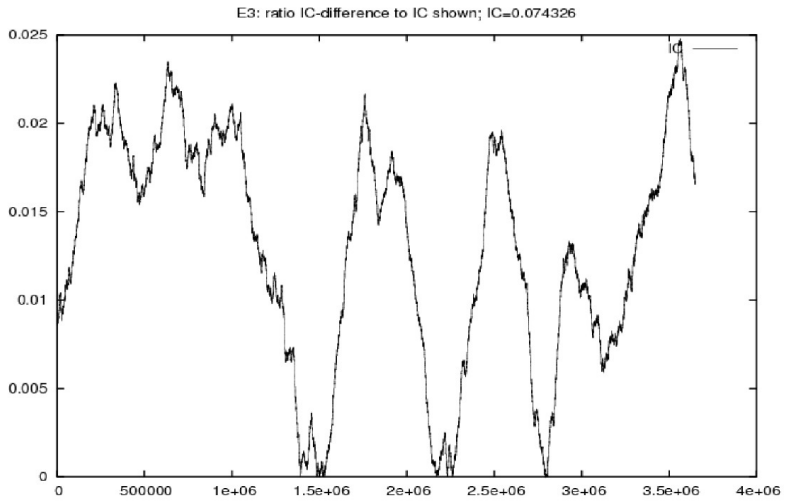
- Zwei Varianten:
  - Typ I: abwechslungsreiche, interessante Bereiche
  - Typ II: für das ganze Korpus repräsentative Bereiche
- Typ I: entropiebasiert
  - erstelle für das ganze Korpus LZMW78-Wörterbuch
  - komprimiere Korpus mit dem Wörterbuch
  - schlecht komprimierbare Bereiche haben höchsten Entropiegehalt -> höchste Informationsdichte
- Typ II: basiert auf Koinzidenzindex (KI)
  - Zahl, wie oft ein Zeichen an derselben Stelle in zwei Texten erscheint, geteilt durch die mögliche Gesamtzahl solcher Paare -> typischer Wert je Sprache
  - Berechnung des KI für das gesamte Korpus und Ausschnitte, kleinere Differenz ist besser

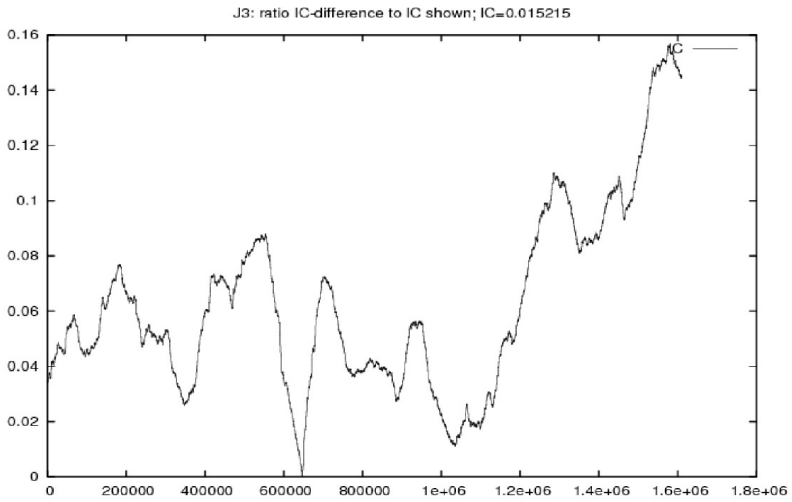
# Exzerpieren von Korpora: Typ I - Beispiel H1

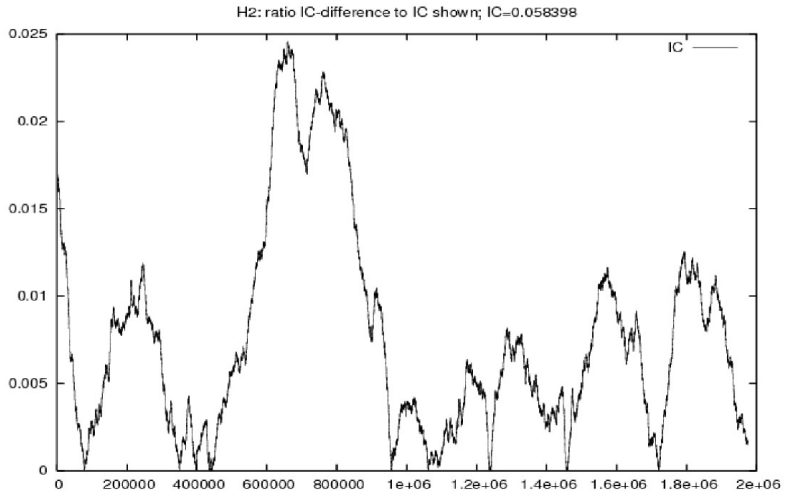


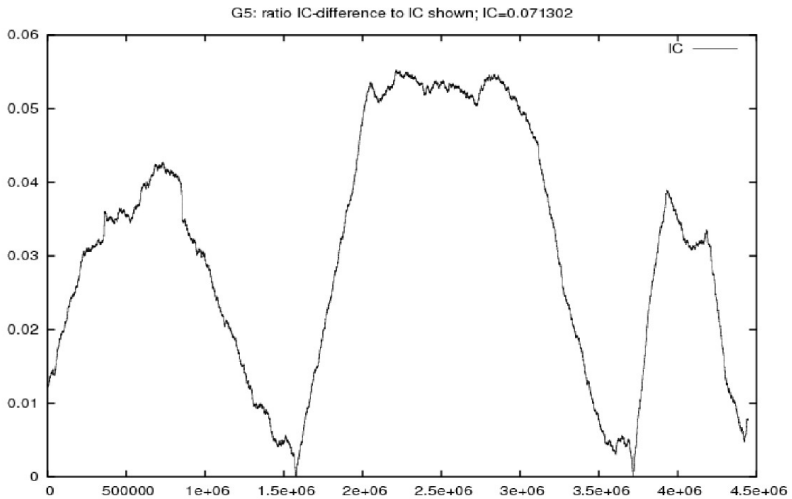


# Exzerpieren von Korpora: Typ II - Beispiele E3, J3, H2, G5



# Detektion syntaktischer Trenner

- wichtige Voraussetzung für weitere Verfahren
- Methoden und Verfahren:
  - Zeichenhäufigkeit
  - Wiederholungen
  - Pangramme
  - Kompression und LCS
  - Aligner

# Zeichenhäufigkeit

- Annahme: Zeichen, die am häufigsten vorkommen, sind am ehesten syntaktische Trenner.
- Ergebnisse:
  - Englisch: “ etao”, “ etoa”, “ etoh”
  - Deutsch: “ enir”, “ enir”, “ enir”, “ enir”, “ enir”
  - Hebräisch: “ YWMH”, “ YFWA”
  - Japanisch: “ 、の。たし”, “ 、のたしい”, “ 、の。たし”, “ のいたしに”, “ の、。たに”
- J4 (kleines Korpus) und J5 Spezialfälle (Zeitungsarchiv)
- Ergebnisse insgesamt recht gut, aber nicht perfekt => Zusatzkriterium

# Wiederholungen

- Idee: syntaktische Separatoren müssen nicht wiederholt werden => aufeinanderfolgende Wiederholungen von Zeichen vermindern die Wahrscheinlichkeit, dass es einer ist
- Problem: “...” => Einzelzeichen erhalten spezielle Bedeutung bei Wiederholungen
  => nur Zusatzkriterium
- Anmerkung: “...” hat eine semantische Bedeutung, daher könnte man es evtl. ignorieren; oder umgekehrt: Wurde “.” als syntaktischer Separator identifiziert, dann könnte eine besondere Bedeutung bei Wiederholungen angenommen werden.

# Pangramme

- Pangramm: Zeichenkette, in dem kein Zeichen zweimal vorkommt
- Idee: Kombination von Zeichenhäufigkeit und Wiederholungen => zerlege Text an Pangramm-grenzen, zähle mit, welches Zeichen dies am häufigsten verursacht
- Beispiel: “... Now| the| earth| was| formles|s and| empty.| Darknes|s wa|s on| the| surface ...”
- Ergebnisse:
  - Englisch: “ etol”, “ etio”, “ elts”
  - Deutsch: “ enrt”, “ entr”, “ entl”, “ entr”, “ enst”
  - Hebräisch: “ MHWT”, “ FYMH”
  - Japanisch: “ の、しいた”, “ の、たしい”, “ の、したい”, “ のい…した”, “ の、にーた”

# Kompression und LCS

- Idee:
  - Aufbau eines Wörterbuchs mittels LZMW78
  - LCS der Einträge des Wörterbuchs => da Wortform-grenzen nicht eingehalten werden, werden sich viele syntaktische Trenner in den Ergebnissen befinden
  - segmentiere den Text mittels der vermuteten syntaktischen Trenner
  - komprimiere segmentierten Text, bewerte nach
    - < Zeichen im Wörterbuch
    - > (genutzte Einträge)/(Gesamtzahl Einträge)
    - > Durchschnittslänge genutzter Einträge
    - < Zahl der Wiederholungen
- Ergebnisse überall sehr gut, z.B.: E3: “ eao”; G3 “ eai”; H1: “ WY)”; H2: “ FAT”; J3: “ 、し。に”

# Aligner

- Idee: Zähle Anzahl und Länge gleicher Suffixe und Präfixe ausgehend von einem vermuteten syntaktischen Trenner
- richtet Text an Trennzeichen aus => “Aligner”
- Besonderheit: obige Berechnungen werden von unten wie von oben ausgeführt und addiert:
  - von oben: “**Au**to”, “**Au**sfahrt”, “**Aus**sehen”
  - von unten: “**Aus**sehen”, “**Aus**fahrt”, “**Au**to”
- Ergebnisse: E3: “ ehta”; G3: “ enid”; H1: “ YW)M”; H2: “ FYWA”; J3: “ 、。のたし“
  => überall sehr gute Ergebnisse

# Textsegmentierung

- benötigt Segmentierung mittels wenigstens eines syntaktischen Trenners
- Methoden und Verfahren:
  - Suffix- und Präfixdetektion
  - Kompositazerlegung
  - Palindrome
  - statistische Verfahren

# Präfix- und Suffixdetektion

- Verfahren:
  - zählt die Zahl möglicher Präfixe und Suffixe
  - nur die, die bei der Häufigkeit ein Meta-Rating von 50% erzielen, werden gewählt
  - Zusatz: Wahl der Wortformen, die mindestens ein Suffix und ein Präfix aus der Ergebnisliste enthalten => Entfernung der Affixe gibt mit hoher Wahrscheinlichkeit natürliche Worte (ohne “.”, “,”, “:” etc.)
- Beispiele:
  - G3: Pr: “be”, “ge”, “ver”; Suf: “,”, “.”; WF: “bot”, “gebe”, “geben”, “komm”, “schrieb”
  - J3: Pr: “ 「”; Suf: “ は”, “ も”; WF: “ あなたがた”, “ わたし”, “ 人々”, “ 神様”

# Präfix- und Suffixdetektion

- Wiederholungen möglich, Beispiel J3:
  Pr: “ イエス”, “ キリスト様”, “ 主”, “ 彼ら”, “ 神”, “ 私たち”; Suf: “ て”; WF: “ の”
  => zeigt: nicht immer werden Affixe oder WF als solche detektiert, die Ergebnisse sind aber akzeptabel und in weiteren Schritten verwendbar
- Beispiel J3:
  “ モーセ | の | 五書旧約聖書 | の | 初めから五冊目までを指す名称で | 、 | 旧約聖書 | の | 中心的部分を占め | て | います。ほかに | 「 | 法律 | の | 書」とも呼ばれます。 | 神様 | はイスラエル民族を選び | 、 | 彼ら | をとおし | て | 、 | 全世界に祝福をもたらすことを約束なさいました。”

# Kompositazerlegung

- weitere Zerlegung, Reduktion Wörterbuch
- Verfahren:
  - erstelle Wortformenliste aus segmentiertem Text
  - zerlege Einträge des Liste, wenn ihre Teile einzeln nachgewiesen werden können
  - bei mehreren Zerlegungsmöglichkeiten wähle diejenige, die möglichst gleich lange und möglichst wenige Strings produziert (Bsp.: “where|upon” statt “where|up|on”)
- Beispiele:
  - “As|he|rim”, “no|on”, “to|get|her”
  - “Arm|er”, “Blut|schuld”, “Blut|vergieße|rinnen”, “darnieder|gelegt”, “hoch|geh|alten”
  - “ おまえ | たちは | もう | 二度と“ ,“ 熱心 | 真心から“

# Palindrome

- Zeichenfolgen, die von hinten wie von vorne gelesen identisch sind (Bsp. “nun”, “noon”, “ まま” ; “enne”, “ele”, “esse”, “s s”)
- Ideen:
  - (partielle) Palindrome können eingesetzt werden, um die Zerlegung üblicher Affixe (“esse”) zu verhindern
  - in einem nichtsegmentierten Text sind in partiellen Palindromen häufig syntaktische Separatoren zu finden (“s s”, “t t”, “o o”)
- Ergebnis: nicht universell (kaum Palindrome in Japanisch) , nicht nur gut
  - manchmal gut: “Interesse|n” statt “Interess|en”
  - verhindert manchmal gute Zerlegungen: “ler|nen”

# Statistische Verfahren

- drei verschiedene Verfahren:
  - Bigramme
  - Häufigkeitsstatistiken von Strings beliebiger Länge
  - Strings fester Länge mit einem Meta-Rating >50%
- Bigramme:
  - Idee: Zerlegung des Textes in häufig vorkommende Einheiten, z.B. Silben oder Wörter (J.)
  - Ablauf: Ist in einem String “abc” das Bigramm von *a* und *b* größer als das von *b* und *c*, so werden sie zusammengelassen; ansonsten werden sie getrennt.
    - “Am| |An|f|ang| |s|ch|uf| |Gott| |Hi|m|mel| |und| |Er|de.”
    - “ モーセの | 五書 | 旧約 | 聖書の | 初め | から | 五 | 冊目 | まで | を | 指す | 名 | 称 | で、 | 旧約 | 聖書の | 中心的 | 部 | 分を | 占 | め | てい | ま | す。”
      => besser für Sprachen mit großer Zeichenanzahl

# Statistische Verfahren

- Häufigkeitsstatistiken von Strings beliebiger Länge:
  - Erweiterung des Bigrammansatzes auf beliebig lange Zeichenketten, an jeder einzelnen Stelle on-the-fly berechnet
  - Änderung: kein Bigramm mehr, sondern Häufigkeit
  - Ergebnisse, exemplarisch an E3:
    “...|In |the |beginning |God |created |the |he|avens |and |the |earth. |Now |the |earth |was |formless |and |empty. |Da|rk|ne|ss |was |on |the |surface |of |the |deep. |God's |Spirit |was |ho|ver|ing |over ...”
    => kaum Auswirkungen, bei anderen Sprachen vergleichbar

# Statistische Verfahren

- Strings fester Länge mit einem Meta-Rating >50%
  - Idee: häufig vorkommende Strings sind wahrscheinlich Wortformen => können für weitere Segmentierung verwendet werden
  - Bewertung der Häufigkeit: MR >50%
  - Beispiele:
    - E3: (2) “ a”, “ t”, “he”, “th”; (3) “ th”, “he “, “the”
    - G3: (2) “ d”, “ ,”, “ch”, “en”; (3) “ de”, “ un”, “und”
    - H1: (2) “ )”, “ H”, “ W”, “T “; (3) “ . “, “ WY”, “YM “
    - H2: (2) “ )”, “ H”, “OW”, “UW”; (3) “ . “, “ HA”, “ WA”
    - J3: (2) “ から”, “ した”, “ です”, “ す。”, “ まし”; (3) “ した。”, “ です。”, ” ました”
      => auch wenn länger nur wirklich gut für Japanisch; Minimumlänge ist im Voraus ja nicht bekannt

# Fazit

- Ziel der Diplomarbeit wurde erreicht
- ohne Vorwissen ist nun möglich:
  - Exzerpieren
  - Detektion syntaktischer Separatoren
  - Detektion von Suffixen, Präfixen und Komposita
  - Zerlegung von Strings in kleinere Einheiten
- entwickelte Algorithmen funktionieren mit allen untersuchten Sprachen
- zeigt: viel Information über die Sprache lässt sich von der Ausdrucksebene her erschließen
- automatische Zerlegung bietet teilweise überraschende Resultate (vgl. Komposita)
- Verzicht auf absolute Thresholds möglich

# Ausblick

weitere mögliche Aufgaben:

- Überprüfung mit weiteren Sprachen und anderen Arten von Korpora
- Reihenfolge und Häufigkeit der Anwendung der Verfahren
- Identifikation von Autoren, unterschiedliche Reihenfolge der Erstellung eines Textes (Stil)
- Pangramme (u. Palindrome) als Bewertungsfunktion: je weniger lange Pangramme existieren, desto besser
- weitere Analyse von Verfahren, die mit diesem Feld recht wenig zu tun haben (vgl. Koinzidenzindex, LCS)

# Diplomarbeit

Quellenangaben, weitere Ergebnisse und Details zu Sprachen, Techniken, Konzepten und Implementationen der Algorithmen in:

“Unsupervised Analysis and Segmentation of Strings of Natural Languages on the Expression Level”

(Originaltitel der Diplomarbeit)