Tutorium der DGfS-Sektion Computerlinguistik

KORPUSLINGUISTIK MIT ONLINE-RESSOURCEN

EINE INTERAKTIVE EINFÜHRUNG FÜR LINGUISTEN

Satzalignment und Übersetzungskandidaten

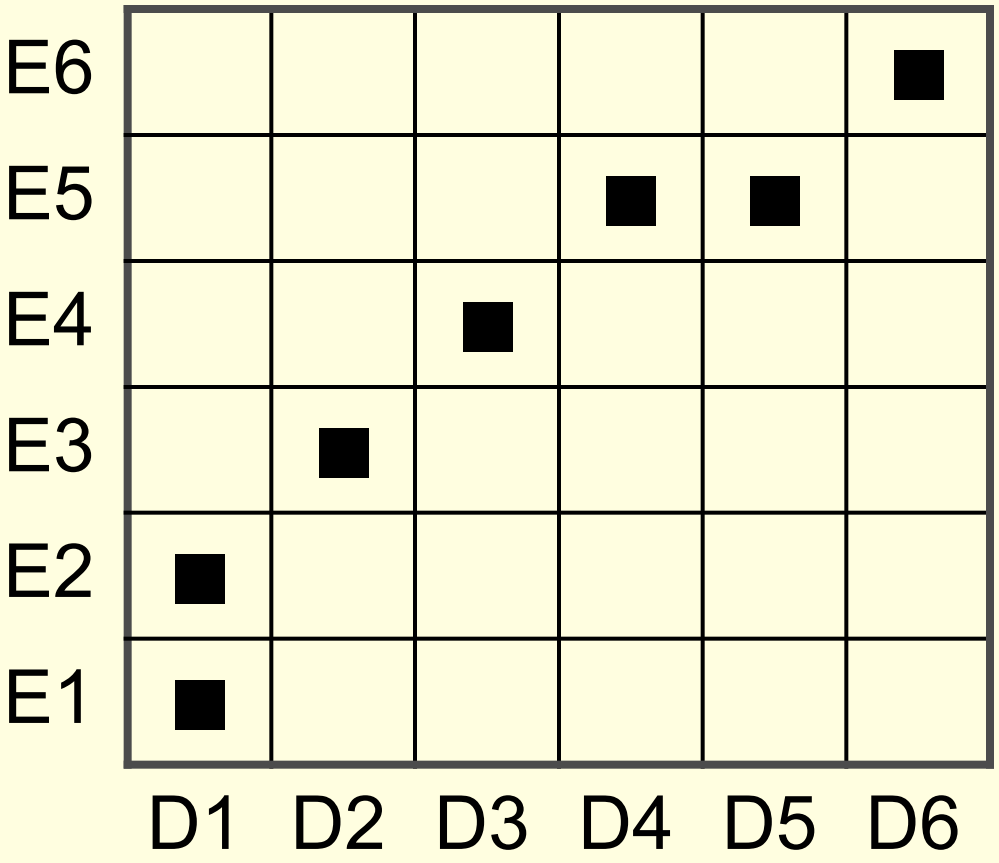
Stefanie Dipper
Stefan Evert
Heike Zinsmeister

München, 28.01.2011

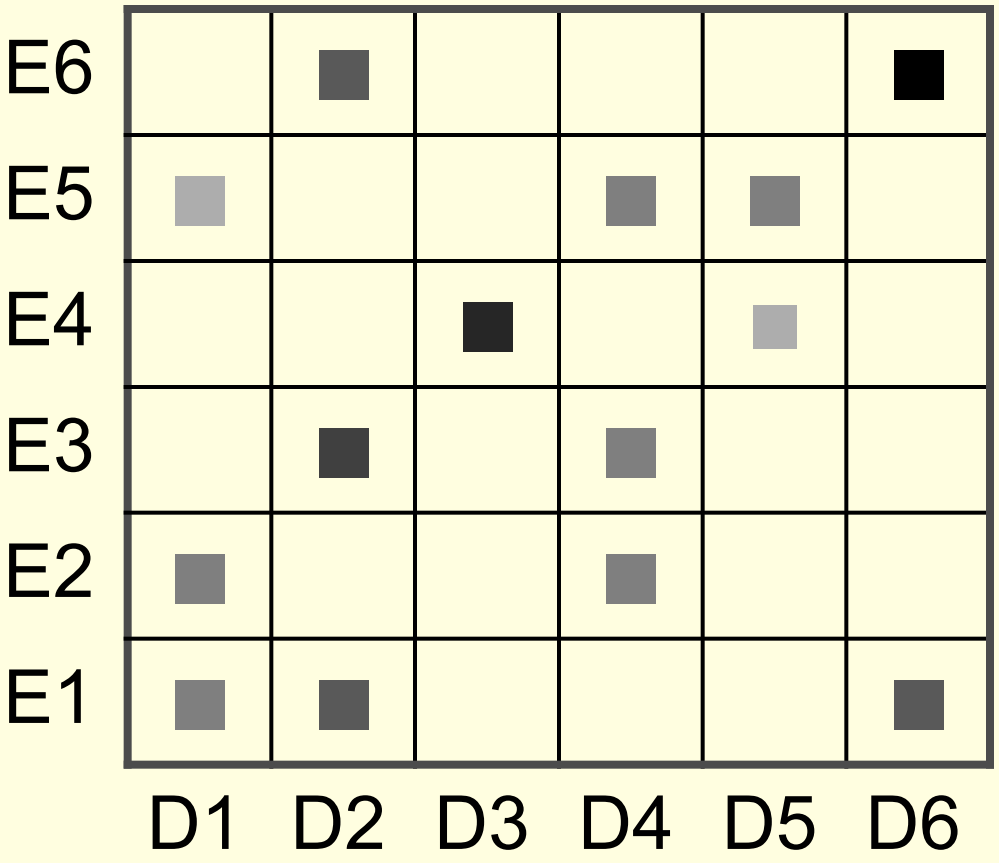
Satzalignment & Bitext

Das stört mich keineswegs, ich halte das für eine gute Initiative, aber wiederum ist Europa nicht zur Stelle.	That is no problem for me.
	I think it is a good initiative, but again Europe is absent.
Es darf nicht wieder geschehen!	It should not happen again, Mr President.
Meine Fraktion verlangt, daß die italienische Präsidentschaft hier vor uns erklärt, welche Rolle sie spielt.	My Group wants the Italian presidency to come here and explain what its role is.
Herr Präsident, liebe Kolleginnen und Kollegen!	Mr President, ladies and gentlemen, I think it is important that we should discuss the situation in the Middle East this week.
Ich halte es für wichtig, daß wir diese Woche über die Situation im Nahen Osten reden.	
Darin sind wir uns alle einig.	We all agree on that.

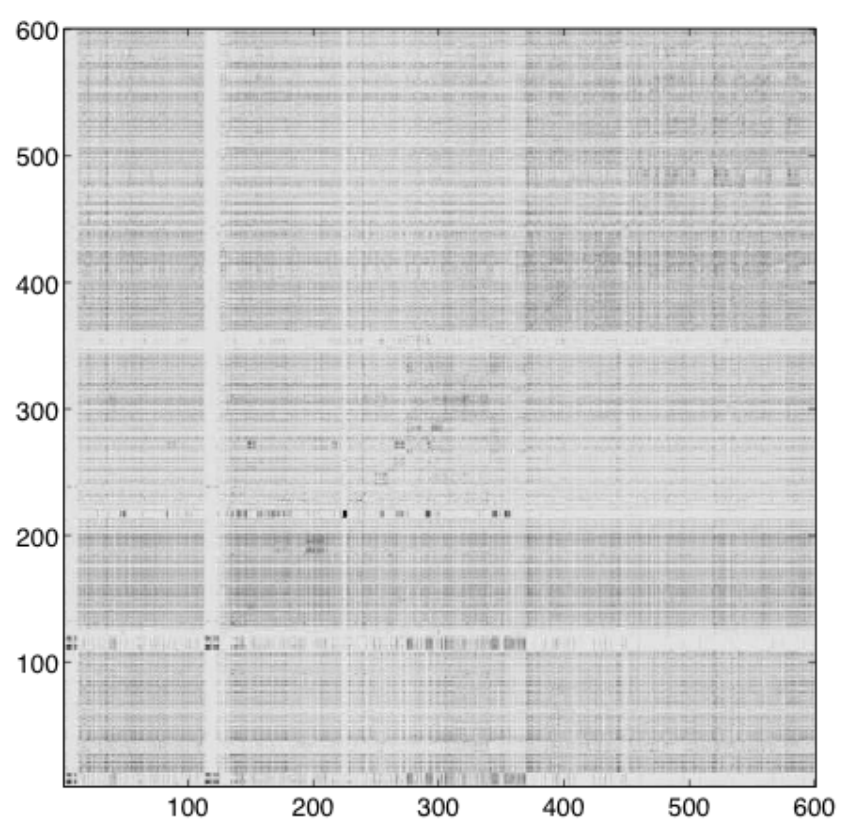
Satzalignment & Bitext

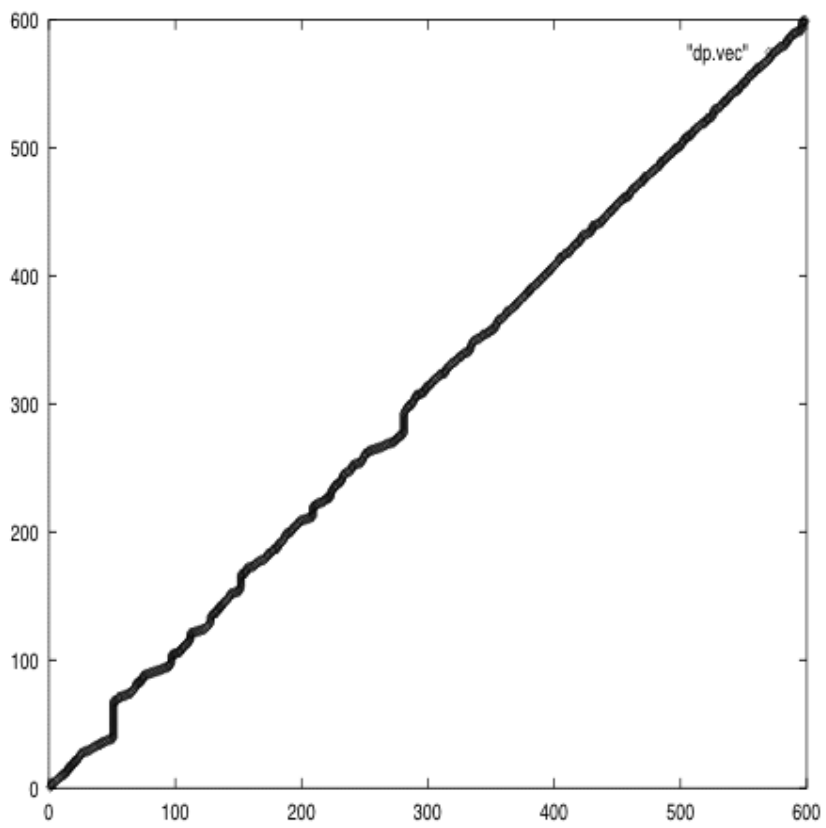


Alignment als Ähnlichkeitssuche

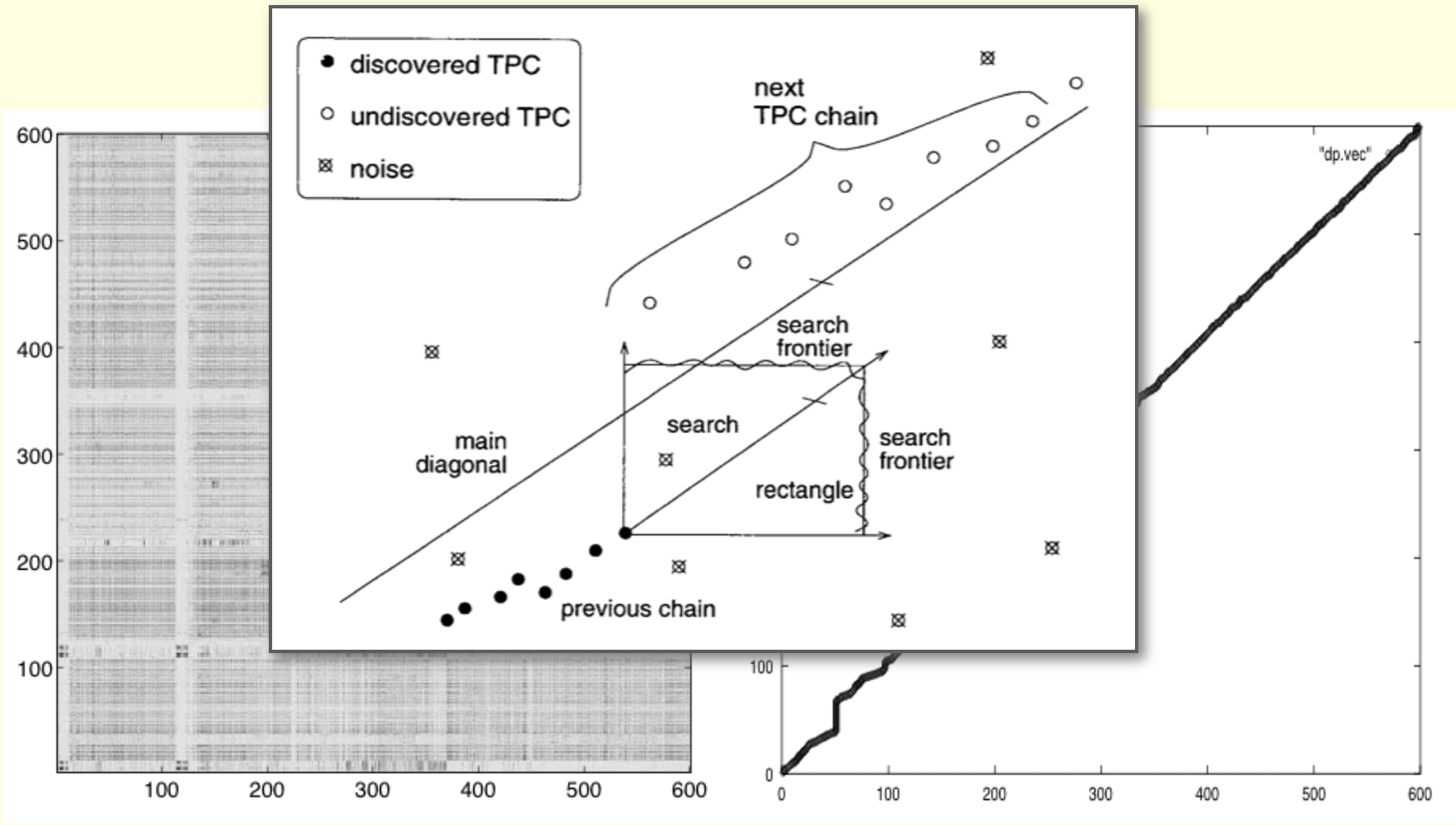


Alignment als Ähnlichkeitssuche





Alignment als Ähnlichkeitssuche



Übersetzungskandidaten

- Einfaches Verfahren zur Suche nach Übersetzungskandidaten:
  - englisches Übersetzungsäquivalent eines deutschen Wortes sollte relativ häufig in alignierten Sätzen vorkommen
  - Übersetzungskandidaten als Kollokationen
  - verschiedene Assoziationsmaße (AM): unser System verwendet Dice-Koeffizient

Übersetzungskandidaten

- *Gesetz* kommt in $f_1 = 4.849$ Sätzen vor
- law kommt in $f_2 = 18.117$ Sätzen vor
- $f = 2.539$ davon sind alignierte Satzpaare
  - bei zwei völlig unterschiedlichen Wörtern würden wir ca. 50 Satzpaare erwarten
  - z. B. *Gesetz* und *lady* in 42 Satzpaaren
- Dice-Koeffizient: $2 * f / (f_1 + f_2) = 0.22$
  - 0.22 ist relativ großer Wert (Bereich: 0…1)