

Tesla - ein Labor für Computerlinguisten

Jürgen Hermes & Stephan Schwiebert

Sprachliche Informationsverarbeitung
Institut für Linguistik
Universität zu Köln

01.06.2012



Anwendungsfall 1

Anwendungsfall 2

Arbeiten mit Tesla

Fazit und Diskussion

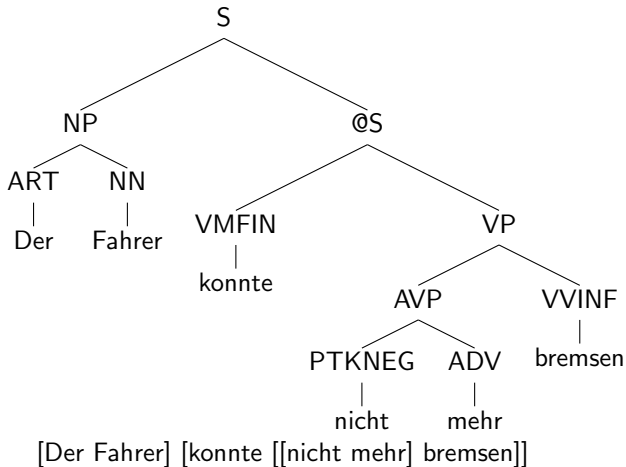


Anwendungsfall 1:

Extraktion syntaktischer Struktur aus unstrukturierten natürlichsprachlichen Texten durch Alignment-Verfahren, bzw.:
Inwieweit lassen sich implizit vorliegende grammatische Strukturen durch ein unüberwachtes Verfahren explizit machen?

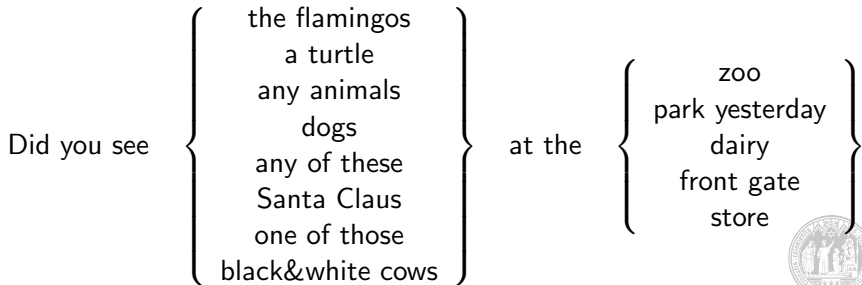


Syntaktische Analyse: Trainierter Parser

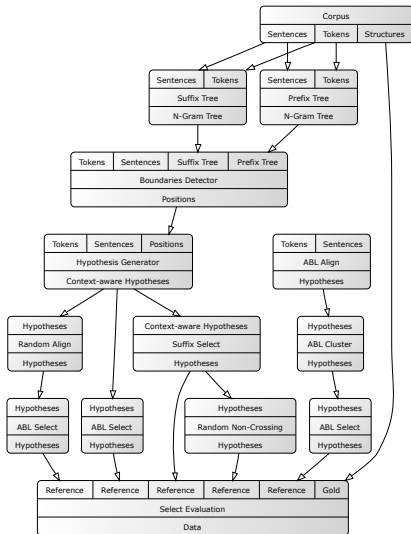


Alignment: Idee

- ▶ paarweiser Vergleich aller Sätze
- ▶ Hypothesenbildung anhand identischer (Wort-)Sequenzen
- ▶ Auflösung widersprüchlicher Hypothesen



Analyse in Tesla



Variation in

- ▶ Wahl der Korpora
- ▶ Art der Vorverarbeitung
- ▶ Hypothesen-Generierung
- ▶ Hypothesen-Selektion

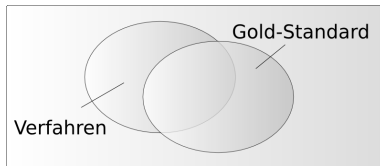
Auswertung anhand

- ▶ Gold-Standard
- ▶ Berkeley-Parser
- ▶ linguistischer Strategie
- ▶ Zufalls-Verfahren



Vergleich mit Berkeley Parser: *Unlabeled Bracket F-Score*

	TB-D/S	TB-E/S	TB-D/Z	BNC-G	CHILDES (3-6)
Right	55,73	40,72	38,65	35,2	61,24
Align (ST)	43,03	36,55	19,19	21,21	52,58
Align (ABL)	23,72	28,67	14,69	14,98	39,29
Random	17,95	18,34	9,81	10,69	33,42



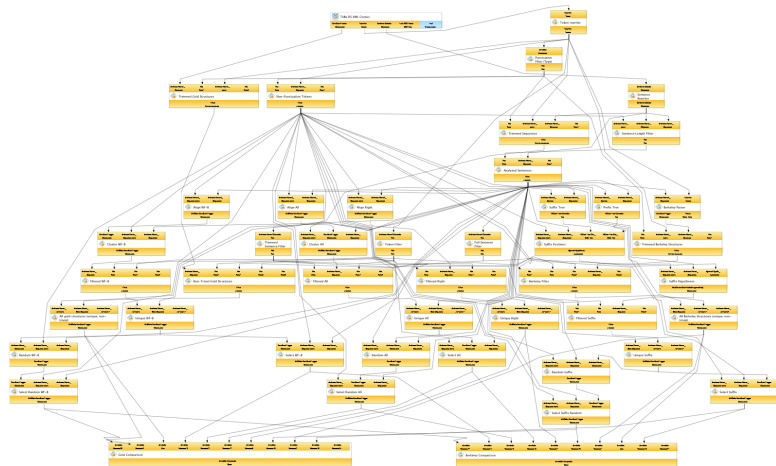
► Precision: $\frac{|V \cap G|}{|V|}$

► Recall: $\frac{|V \cap G|}{|G|}$

► F-Score: $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$



Beispiel-Versuchsaufbau



Anwendungsfall 2:

Analyse der Eigenschaften des Textes aus dem Voynich Manuskript;
Suche nach einer Chiffre mit ähnlichen Eigenschaften; Nutzung der
Wortstruktur als denkbarer Weg zur Entschlüsselung oder:
Wie operationalisiert man Schrägheit?



Texte als Sequenzen diskreter Einheiten

Georg Trakl

De profundis

Es ist ein Stoppelfeld, in das ein schwarzer Regen fällt.
Es ist ein brauner Baum, der einsam dasteht.
Es ist ein Züchelhund, der leere Hütten umkreist.
Wie trauig dieser Abend.

Am Weiler vorbei
Sammelt die sanfte Waie noch spärliche Ähren ein.
Ihre Augen werden rund und goldig in der Dämmerung
Und ihr Schoß hart des häuslichen Bräutigams.

Bei der Heimkehr
Finden die Hürten den süßen Leib
Verwest im Dornenbusch.

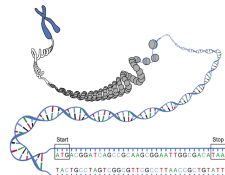
Ein Schatten bin ich ferne finsternen Dürfern.
Gottes Schweigen
Trank ich aus dem Brinnen des Hains.

Auf meine Stirne tritt kaltes Metall
Spinnen suchen mein Herz.
Es ist ein Licht, das in meinem Mund erlöschet

Nachts fand ich mich auf einer Heide,
Starrend von Urart und Staub der Sterne.
Im Haezelgebirch
Klangen wieder kristallne Engel.



Es ist ein Stoppelfeld, in das ein schwarzer Regen fällt.
Es ist ein brauner Baum, der einsam dasteht.
Es ist ein Züchelhund, der leere Hütten umkreist.
Wie trauig dieser Abend.



ab	04	af	c2	bf	a3	81	ec	00	01	06	00	31	e8	88	0c
bc	fe	c1	75	29	31	c0	ba	af	be	ad	de	02	04	0c	00
de	c1	ca	08	8a	1c	0c	8a	3c	04	88	1c	04	88	3c	0c
fe	c1	75	e8	a9	3c	00	00	00	89	e3	81	c3	04	00	00
00	0c	58	3d	41	41	41	41	75	43	58	3d	42	42	42	42
75	3b	5a	89	d1	89	a6	89	df	29	cf	f3	a4	89	de	89
41	89	df	29	cf	31	c0	31	db	31	d2	fe	c0	02	1c	06
8a	14	06	8a	34	1e	89	34	06	88	14	1e	00	22	30	16
8a	1c	16	8a	17	30	de	89	17	47	49	75	de	31	de	89
49	fe	c0	cd	80	90	90	e8	3d	ff	ff	41	41	41	41	41



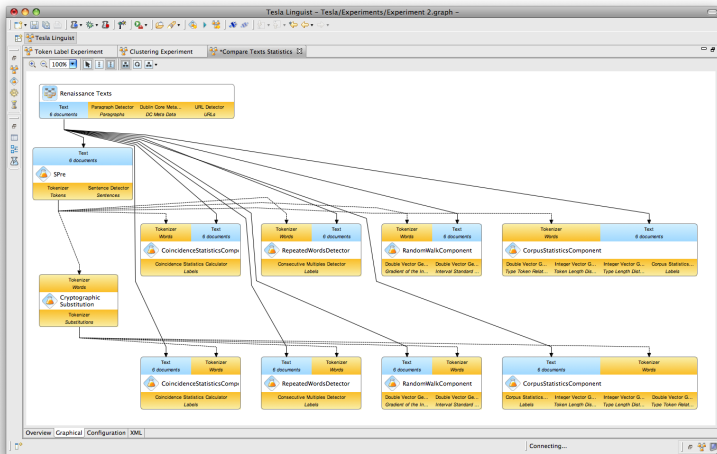
Reproduzierbarkeit

Reproduzierbarkeit als zentrale Forderung für Wissenschaftlichkeit

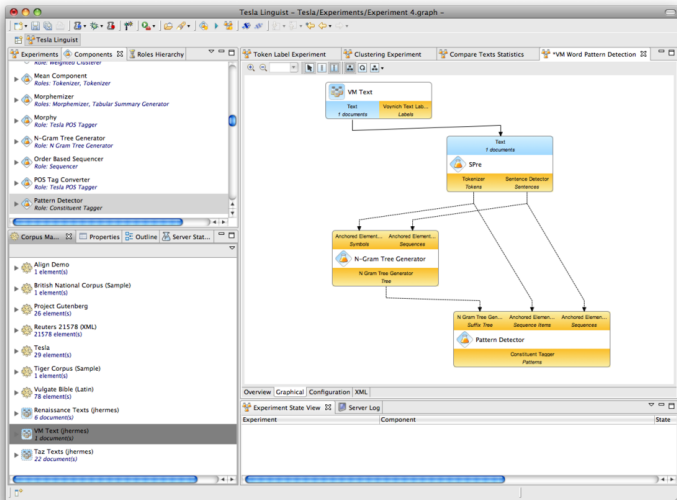
- ▶ Yale Roundtable on Data and Code Sharing (2010):
“Reproducibility should become the responsibility of a scientific community”
- ▶ Ted Pedersen (2008): “We publish page after page of experimental results where apparently small differences determine the perceived value of the work. [...] Our inability to reproduce results leads to a debilitating paradox, where we as reviewers and readers accept highly empirical results on faith.”



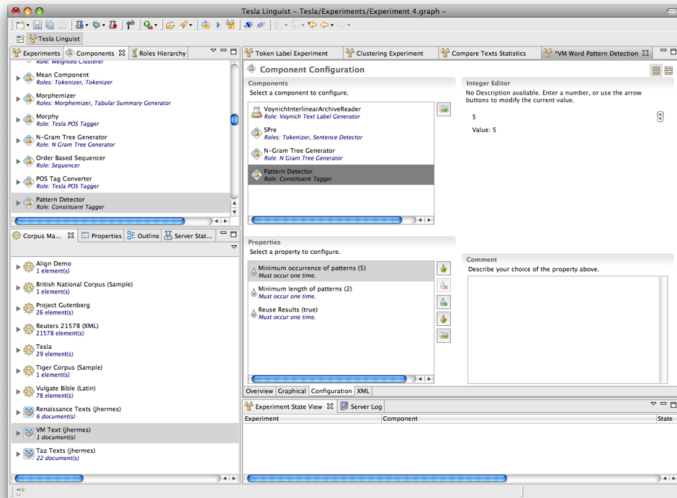
Umsetzung in Tesla



Tesla - Workflows für die Textprozessierung



Tesla - Workflows für die Textprozessierung



Tesla - Workflows für die Textprozessierung

The screenshot displays the Tesla Linguist application window, titled "Tesla Linguist - Evaluation view of experiment VM Word Pattern Detection -". The interface is divided into several panes:

- Left Pane:** A list of experiment tabs including "Token Label Expe...", "Clustering Experi...", "VM Word Pattern D...", and "VM Word Pattern D...". Below these is a large text area containing a complex string of tokens and patterns, with several segments highlighted in yellow.
- Right Pane:** A "Constituent Tagger of Pattern Detector" window. It has two sub-tabs: "Summary" and "Details". The "Details" tab is active, showing a table of properties and their values.

Property	Value
Interfaces	ISingleValueCategory, ILabelledElement
Interval	128513 to 128525
Content	shedy.qokedy
Type ID	-327057111959886531
Role ID	17
Category	2
Label	2

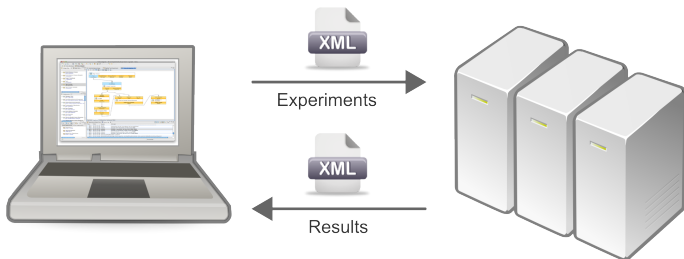
Below the table, there are three more entries for the same properties, each with a different value for the "Interval" and "Content" fields.

The bottom of the window shows a status bar with "Connecting..." and a set of navigation icons.



Tesla-Architektur

- ▶ Tesla Client: Eclipse-basierter *Rich Client*
- ▶ Tesla Server: Built-in oder Standalone



Tesla-Ressourcen

- ▶ **Tesla-Webseite:** <http://tesla.spinfo.uni-koeln.de>
- ▶ **Download:** Windows, Linux, MacOS, jeweils 32 oder 64 bit
- ▶ **Dokumentation** der existierenden Komponenten und Rollen
- ▶ **Tutorials** zur Erläuterung der wichtigsten Funktionen
- ▶ **Lizenz:** *Eclipse Public Licence (EPL)*
- ▶ **Blog:** <http://texperimentales.hypotheses.org>



Tesla in MyExperiment

- ▶ **MyExperiment** (<http://www.myexperiment.org>) ist ein soziales Netz für Wissenschaftler, über das Forschungsobjekte wie Workflows, Daten und Informationen ausgetauscht werden können.
- ▶ Momentan 10 Tesla-Beispiel-Experimente online
- ▶ Vereinfachter Import und Export Tesla $\Leftrightarrow$ MyExperiment in Planung



Tesla-Weiterentwicklung

- ▶ **Anschlussfähigkeit:** Zugriff auf die Ressourcen anderer Frameworks (evtl. TextGrid), Bereitstellung von Diensten für andere Frameworks
- ▶ **Integration neuer Features:** Neue Komponenten, mehr Möglichkeiten der Datenvisualisierung, Verbesserung der IDE, Verbesserung der Benutzerführung
- ▶ **Little Big Data:** Erweiterung des Servers um Clustering-Funktionalität/Ausnutzung bereits vorhandener universitärer Ressourcen (z.B. PC-Pools)



Ende

- ▶ Vielen Dank für eure Aufmerksamkeit!
- ▶ <http://tesla.spinfo.uni-koeln.de>

