

DANTE-Herbsttagung 2013 (02.11.2013, Universität zu Köln)

**Wortlisten:
Voraussetzung für gute Trennmuster**

Referent: B. Sc. cand. geogr. Tobias Wendorff
tobias.wendorff@tu-dortmund.de | @rub.de

Zum Referenten

- bis 2010: Student der Raumplanung (TU Dortmund)
- seit 2010: Student der Geographie (RUB)
- 2007–2013: Hilfskraft der Geographie (TU Dortmund)

- seit 2005: Gewerbe für EDV-Dienstleistungen
- seit 2008: Mitglied des Projekts OpenStreetMap (OSM)

- Spezialisierung auf automatische Datenverarbeitung im Bereich Geodaten (Abgleich, Anreicherung, Darstellung)
- Beratung von Unternehmen beim Umstieg auf OSM und Unterstützung (z.B. Aachener Verkehrsverbund)

rein-vestieren

Fahrer-laubnis

Talent-wässerung

Staub-ecken

Nachteil-zug

Autoren-nen

Wort- und Silbentrennung

- TeX ermittelt optimalen Zeilen- & Seitenumbruch
- im Notfall: Trennung
- manuell: `\hyphenation{ Krüm-mungs-linie }`

Trennungsalgorithmus

- PhD. Franklin Mark Liang (1982)
- Algorithmus basiert auf statistischer Auswertung von Trennmustern
- Leistung maßgeblich von den Mustern abhängig
-

Experimentelle Wortliste „dehyph-exptl“

- Voraussetzung: möglichst komplette Wortliste
- Wörter in allen Flexionsformen
- Einträge der dt. Trennmustermannschaft: 480.000
- Lizenz: LaTeX Project Public License (LPPL)

Organisation der Trennmustermannschaft

- keine Hierarchie, jeder darf alles
 - `http://projekte.dante.de/Trennmuster`
- `git://repo.or.cz/wortliste.git`
- `trennmuster@dante.de`

Inhalte der Wortliste

- geläufiger Schrift- und Sprachbereich in DACH
- Fachsprache, häufige Eigennamen
- Gewichtungen und Kategorisierungen der Trennstellen

Quellen für den Inhalt

- amtliche Regeln zur Rechtschreibung
- frei verfügbare Wortlisten und Referenzen
 - *Google Books*-Korpus,
 - Grimm'sche Wörterbuch etc.

Gewichtung (1)

- an Wortfugen: **Wort=fu-ge**
Glas=per-le, Tisch=ten-nis
- nach Präfixen oder Verbalpartikeln: **Vor | sil-be**
ver | ein-zelt, ab | schrei-ben
- innerhalb eines Worts oder vor einem Suffix:
in-nen, frag-lich

Gewichtung (2)

- unterschiedliche Bindungsstärken:
fern=ab | ge | le-gen fern + (ab + (ge + (le+gen)))
- Beachtung von ungünstigen Trennstellen:
Anal-phabeten, Altbauer-neuerung, beinhalten, Stiefen-keel, Sex-tanten
- doppeldeutige und spezielle Trennungen:
Gastrasse, Ostönnen, Drucker (ck vs. kk), Straße
- diskussionswürdige Trennungen:
In-dus-trie vs. In-dust-rie vs. In-du-strie

Datenstruktur und erzeugte Pattern

- DSV-Datei (Trennung durch Semikolon)
- Felder ≥ 3 beinhalten „Zusatzinformationen“
- Versionierung und Dokumentation im GIT
- Selektion von Haupt- und/oder Nebentrennstellen

Aachen;Aa-chen	.ab2 .abs2 .Ak2
Aachener;Aa-che-ner	.al2 .alt3r
[...]	.alt1s .amt2
Aachens;Aa-chens	.an4 .angs2 .ar2
Aalbeck;Aal=beck	.arm3ac .aro2
Aalbestand;Aal=be stand	.as2 .at2 .auf4
	.aus4 .be2 .bei2
	.ben4 .ber4. berg3a

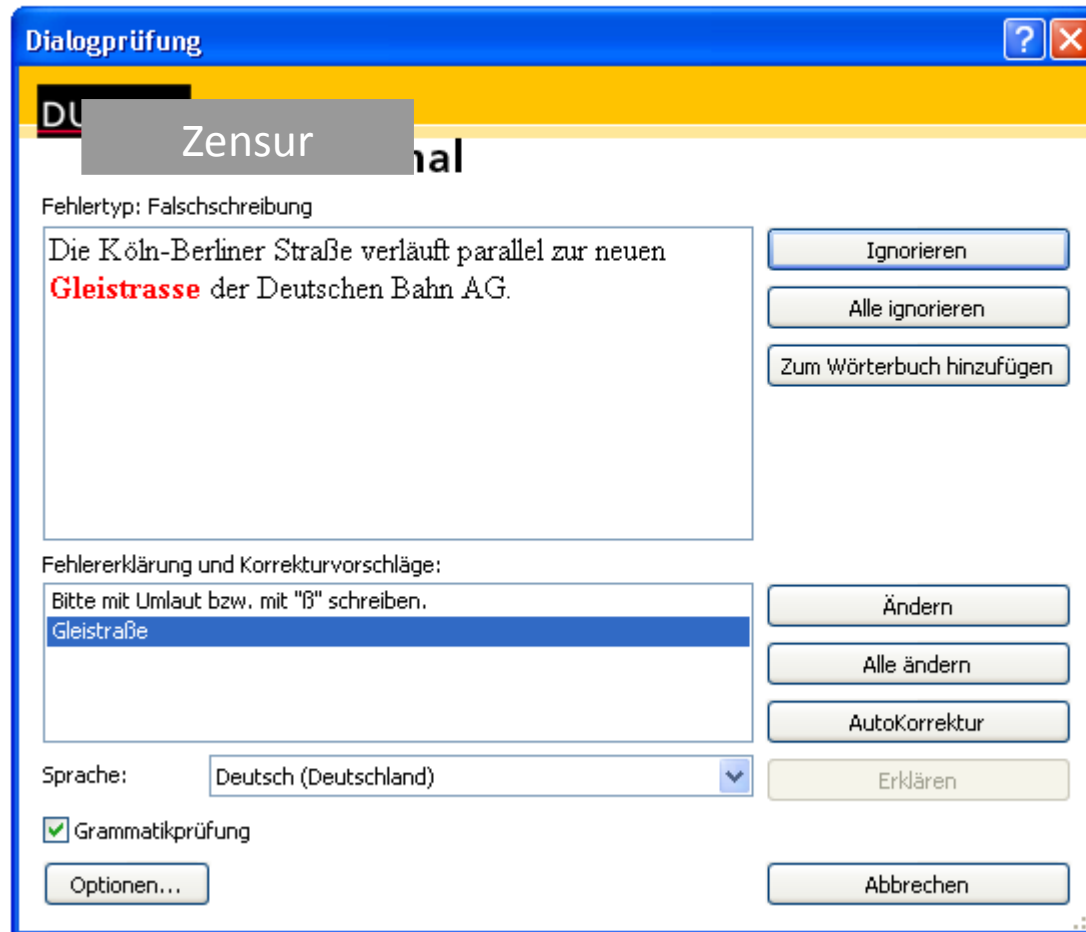
Qualitätssicherung

- hauptsächlich manuell von der Mannschaft
- wenig bis kein Rückfluss aus der Community

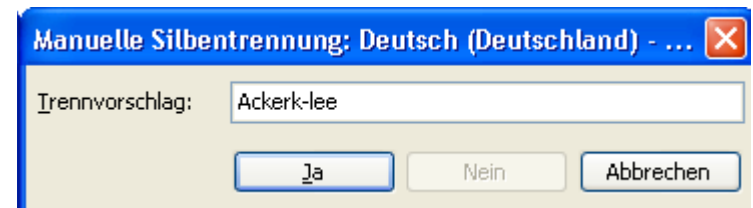
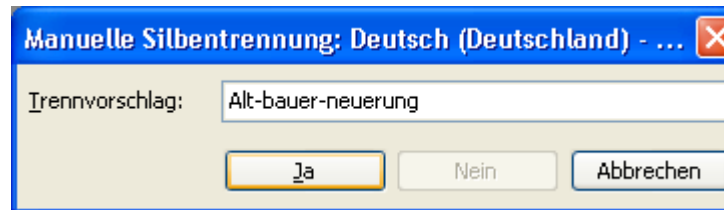
- verschiedene Scripts im Repository (Perl, Python)
 - Überprüfung auf Teilwörter
 - Überprüfung von Prä- und Suffixen

- Abgleich mit vorhandenen Texten (Bücher usw.)
- Abgleich mit kommerziellen Angeboten

Kommerzielle Konkurrenz (1)

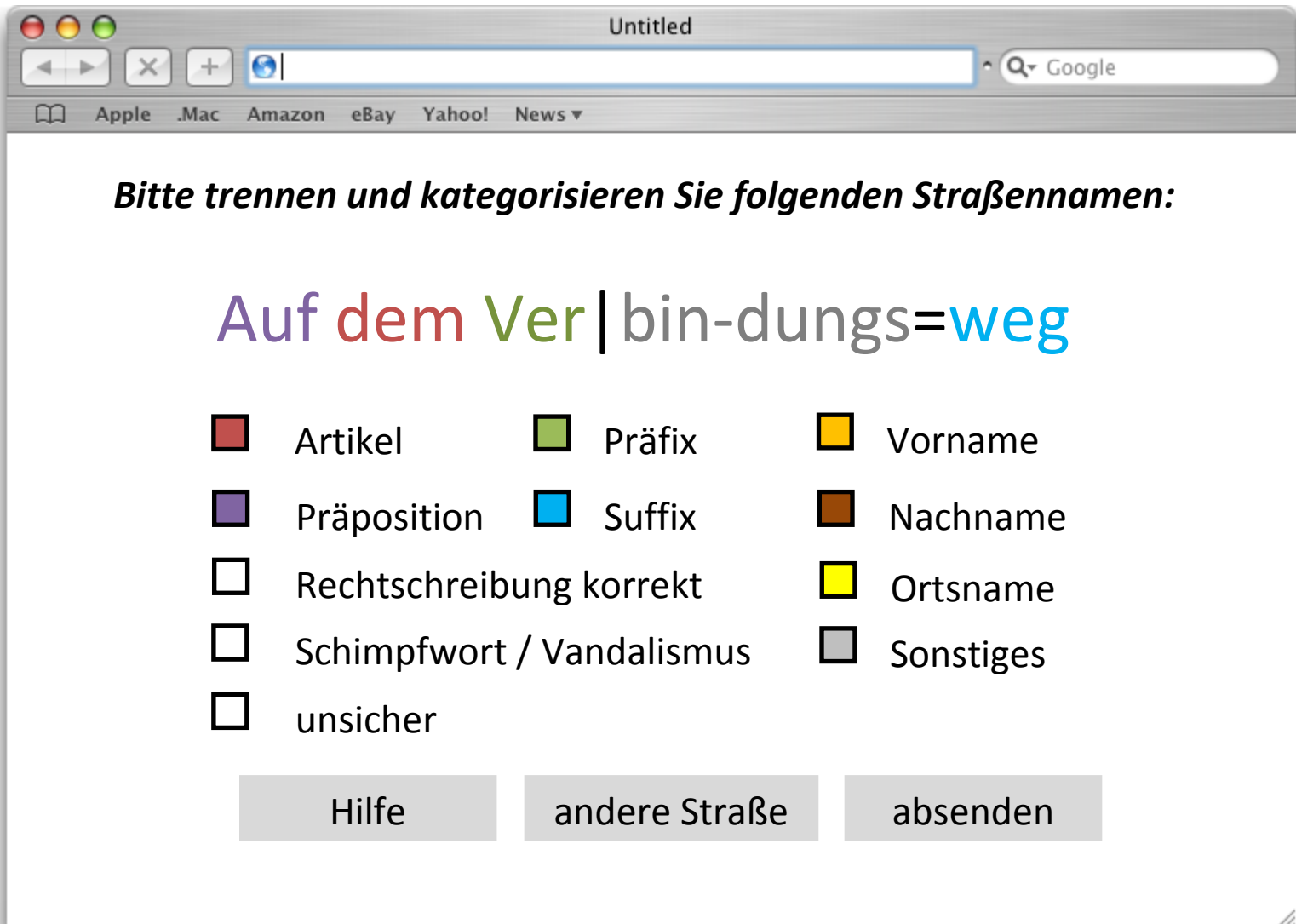


Kommerzielle Konkurrenz (2)



Lösung (1): Crowdsourcing

- Information, dass die Trennmuster auch in anderen Produkten genutzt werden können: Libre-/OpenOffice, Illustrator/InDesign
- Vereinfachung des Kategorisierungsvorgangs durch HTML-basierte Oberfläche
- zufällige, gegenseitige Prüfung des gleichen Wortes (bei drei gleichen Trennungen: Freigabe)



Bitte trennen und kategorisieren Sie folgenden Straßennamen:

Auf dem Ver | bin-dungs=weg

<input checked="" type="checkbox"/>	Artikel	<input checked="" type="checkbox"/>	Präfix	<input checked="" type="checkbox"/>	Vorname
<input checked="" type="checkbox"/>	Präposition	<input checked="" type="checkbox"/>	Suffix	<input checked="" type="checkbox"/>	Nachname
<input type="checkbox"/>	Rechtschreibung korrekt	<input checked="" type="checkbox"/>	Ortsname	<input type="checkbox"/>	Sonstiges
<input type="checkbox"/>	Schimpfwort / Vandalismus				
<input type="checkbox"/>	unsicher				

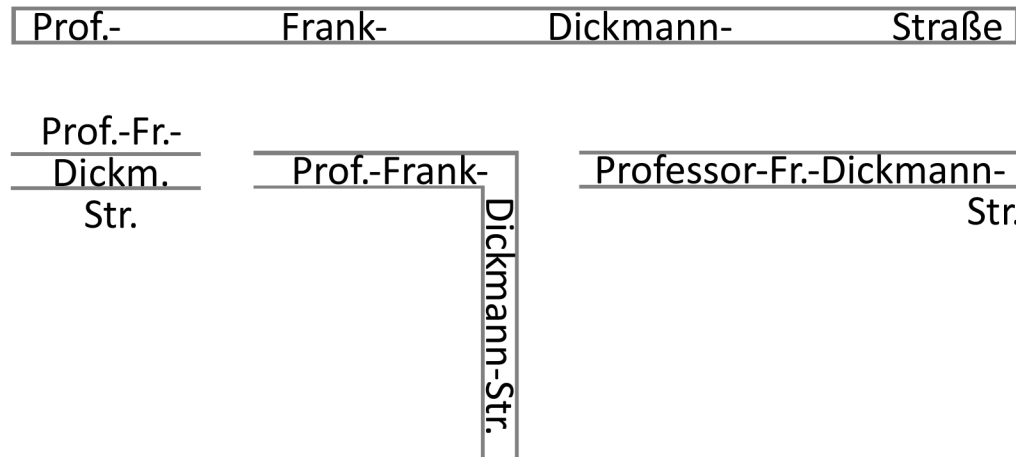
Hilfe andere Straße absenden

Lösung (2): Kooperation mit hunspell (etc.)

- neue Trennmuster helfen die Leistungsfähigkeit des Algorithmus zu erhöhen
- Zugang zu vielen Flexionsformen
- Datenhaltung wird verbessert, manuelle Bearbeitung jedoch erschwert
- Nutzung von Synergien in beiden Projekten

Interdisziplinäre Anwendung: Kartographie

Professor-Frank-Dickmann-Straße



„Endungen“ von Straßennamen

Verkehr

allee, bahn, bogen, brücke, gasse, pfad, plätzchen,
platz, poth, promenade, ring, steige, stieg, stiege,
straße, tangente, umgehung, weg, winkel, ...

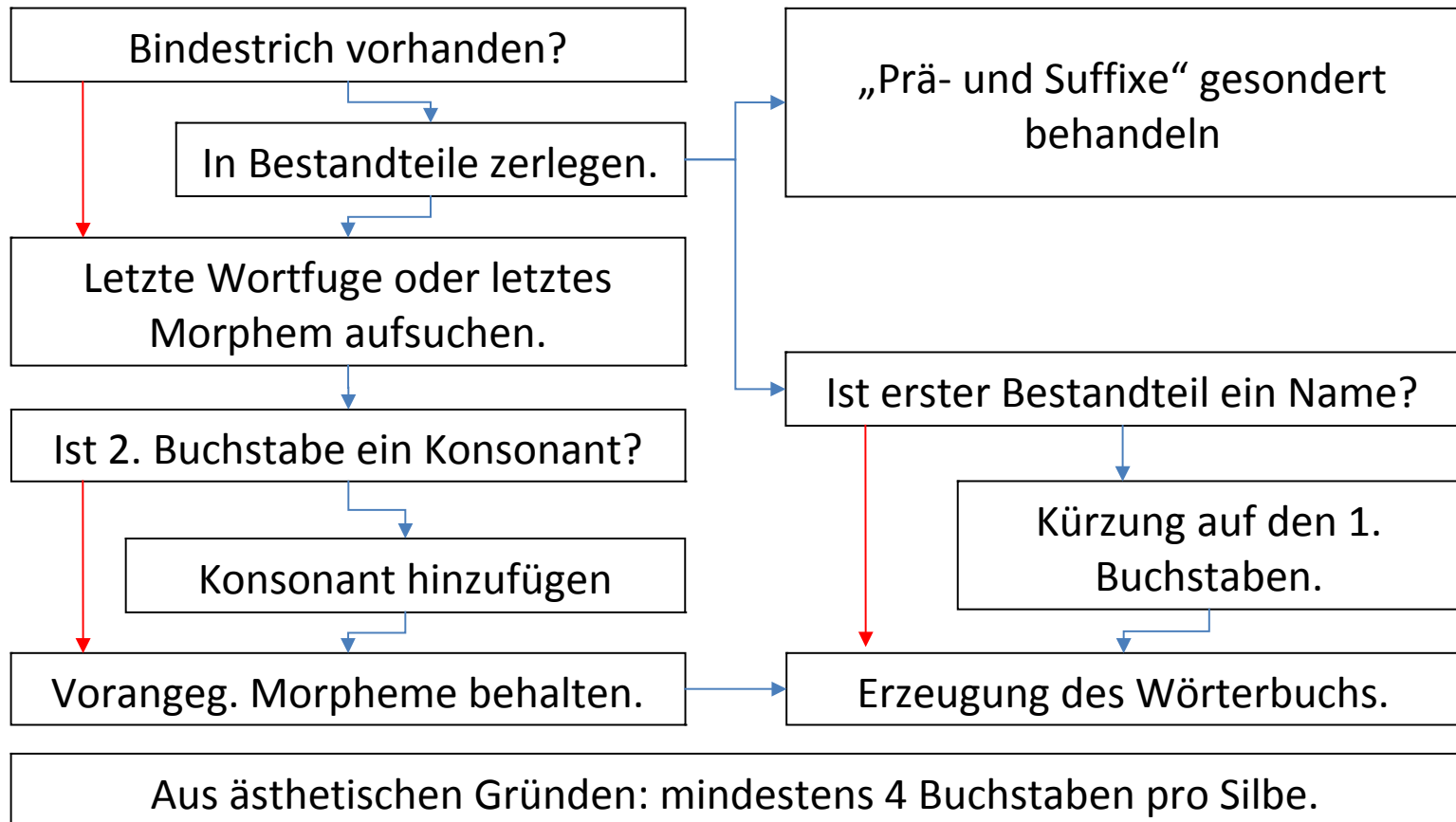
Topographie

acker, bach, bank, baum, beck, becke, berg, blick,
born, brügge, brand, bredde, breite, brink, bruch,
busch, delle, diek, feld, forst, gabel, garten, gatter,
graben, grenze, gut, höhe, höhle, hagen, hain, ...

Abkürzung von „Prä- und Suffixen“

Straße	Str., S.
Stieg	Stg., St.
Weg	Wg., W.
Platz	Pl.
Bahnhof	Bhf., Bf.
Haus	Hs.
Doktor	Dr.
Freiherr	Freih., Frhr., Fhr., Frh.
der/die/das	d.
auf / in	a. / i.
Ost / Ober	O. / Ob.

Festlegung der Abkürzungsregeln (vereinf.)



Fazit

- Algorithmus ist leistungsfähig, es wird jedoch ein hoher Aufwand bei der Erstellung und Wartung der Trennmuster benötigt
- Scripts helfen beim Auffinden der größten Fehler, manuelle Überprüfung jedoch unumgänglich
- Einbindung der Community und Kooperation mit themenverwandten Projekten äußerst sinnvoll

Herzlichen Dank für die Aufmerksamkeit.

tobias.wendorff@tu-dortmund.de | @rub.de

Website der deutschen Trennmustermanschaft:
<http://projekte.dante.de/Trennmuster>