Textdatenbanken

Sommersemester 2020

8. Vorlesung

- Anwendung von parallelem Text -

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Aufgabenstellung

Es gibt für viele Texte Übersetzungen, in einigen günstigen Fällen ist alles frei verfügbar.

Frage: Wie lässt sich daraus ein zweisprachiges Wörterbuch konstruieren?

Oder: Wie lässt sich daraus das Wissen des Übersetzers „rekonstruieren“?

Arbeitsschritte:

Gegeben ist ein (möglicherweise großes) Dokument mit seiner Übersetzung.

Finde immer kleinere zusammengehörige Textteile:

- Absätze
- Sätze
- Wörter

Problem Description

Given:

- certain amounts of sentence-aligned parallel texts

Not available:

- morphology, grammar, semantic etc. information
- string similarity for cognates
- bilingual dictionary

Wanted:

- bilingual dictionaries
- alignment on word level

Broad Picture

- Calculation of translingual statistically significant co-occurrences yields ranked translation candidates

- For alignment, the highest ranked translation candidates that occurr in the sentence pair are linked.

Trans-co-occurrences

Translingual co-occurrences

‘normal‘ co-occurrences:

- Calculaton performed on sentence basis
- Co-occurrents can be found frequently together in sentences

Trans-co-occurrences:

- Calculaton performed on bilingual sentence pairs
- Co-occurrents can be found frequently together in bilingual sentence pairs
- Hypothesis: significant co-occurrences between words of different languages (= trans-co-occurrences) are translation equivalents

Data: Europarl

- Transcriptions of European Parliament, about 1 million sentences per language
- Available for Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portugese, Spanish and Swedish
- Experiments carried out for:
  - Englisch-Danish
  - Englisch-Dutch
  - Englisch-German
  - Englisch-Finnish
  - Englisch-Italian
  - Englisch-Portugese
  - Englisch-Swedish

  (chosen because of dictionary availability)

Example: Gesellschaft@de society@en

Die@de drogenfreie@de **Gesellschaft@de** wird@de es@de aber@de nie@de geben@de .@de *But@en there@en never@en will@en be@en a@en drug-free@en* ***society@en*** *.@en*

Unsere@de **Gesellschaft@de** neigt@de leider@de dazu@de ,@de Gesetze@de zu@de umgehen@de .@de *Unfortunately@en ,@en our@en* ***society@en*** *is@en inclined@en to@en skirt@en round@en the@en law@en .@en*

Zum@de Glück@de kommt@de das@de in@de einer@de demokratischen@de **Gesellschaft@de** selten@de vor@de .@de *Fortunately@en ,@en in@en a@en democratic@en* ***society@en*** *this@en is@en rare@en .@en*

Herr@de Präsident@de !@de Wir@de leben@de in@de einer@de paradoxen@de **Gesellschaft@de** .@de *Mr@en President@en ,@en we@en live@en in@en a@en paradoxical@en* ***society@en*** *.@en*

Ich@de sprach@de vom@de Paradoxon@de unserer@de **Gesellschaft@de** .@de I@en mentioned@en what@en is@en paradoxical@en in@en **society@en** .@en

Zeit@de ist@de Macht@de in@de unserer@de **Gesellschaft@de** .@de *Time@en is@en power@en in@en our@en* ***society@en*** *.@en .*

In all sentence pairs, **Gesellschaft@de** and **society@en** occur together.

Example: top-ranked trans-co-occurrences

Gesellschaft: society@en (12082), social@en (342), our@en (274), in@en (237), societies@en (226), Society@en (187), women@en (183), as@en a@en whole@en (182), of@en our@en (168), open@en society@en (165), democratic@en (159), company@en (137), modern@en (134), children@en (120), values@en (120), economy@en (119), of@en a@en (111), knowledge-based@en (110), European@en (105), civil@en society@en (102)

society: Gesellschaft@de (12082), unserer@de (466), einer@de (379), gesellschaftlichen@de (328), Wissensgesellschaft@de (312), Menschen@de (233), gesellschaftliche@de (219), Frauen@de (213), Zivilgesellschaft@de (179), Gesellschaften@de (173), Informationsgesellschaft@de (161), modernen@de (157), sozialen@de (155), Wirtschaft@de (132), Leben@de (119), Familie@de (118), Gesellschaftsmodell@de (108), demokratischen@de (108), soziale@de (98), Schichten@de (97)

kaum: hardly@en (825), scarcely@en (470), little@en (362), barely@en (278), hardly@en any@en (254), very@en little@en (186), almost@en (88), difficult@en (68), unlikely@en (63), virtually@en (53), scarcely@en any@en (51), impossible@en (47), or@en no@en (40), there@en is@en (38), hardly@en ever@en (37), any@en (32), hardly@en anything@en (32), surprising@en (31), hardly@en a@en (29), hard@en (28)

hardly: kaum@de (825), wohl@de kaum@de (138), schwerlich@de (64), nicht@de (51), verwunderlich@de (43), kann@de (37), wenig@de (37), wundern@de (25), man@de (21), dürfte@de (17), gar@de nicht@de (17), auch@de nicht@de (16), gerade@de (16), überrascht@de (15), fast@de (14), überraschen@de (14), praktisch@de (13), ist@de (12), schlecht@de (12), verwundern@de (12)

Evaluation

- What is the quality of determined translation equivalents?
- Evaluation by comparing results to bilingual dictionaries (freelang) to measure precision
- Method:
  - Only words that are in the dictionary and have automatic translations are taken into account
  - Determine portion of matches in the 3 highest-ranked trans-co-occurrences

Problems:

- Some translations are correct but not found in the dictionary
- Dictionaries are not adopted to domain
- Inflection: Dictionaries contain lemmas -> Prefix matching
- Unknown multiword units

Prefix matching

- Prefix match *prfx(A,B)* of two strings A and B is defined by

$$prfx(\mathrm{A,B}) = \frac{\text{length of common prefix of A and B}}{\max(\text{length(A)}, \text{length(B)})}$$

Examples:

prfx(Herbert, Herberts) = 7/8 = 0.875

prfx(Baustelle, Baugenehmigung)=3/14 = 0.2142

prfx(Häuserkampf, Häuserkämpfe) = 7/12 = 0.5833

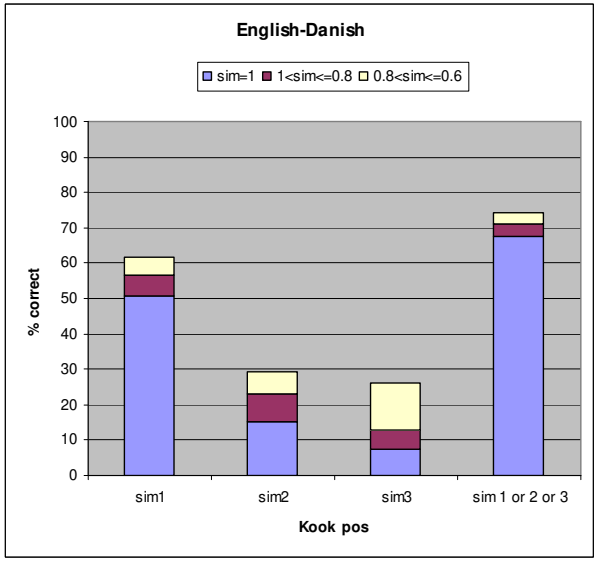
A quite crude measure, but deals more or less with the inflection problem

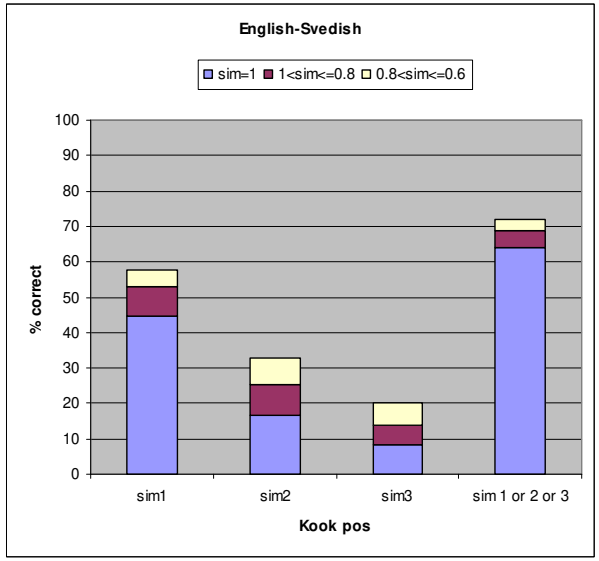
Sample data from en-de

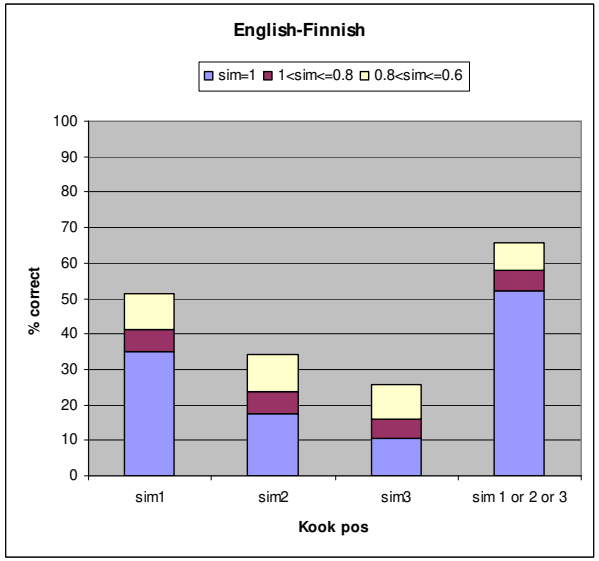
word (en)	co1 (de)	p1	co2 (de)	p2	co3 (de)	p3
absolutely essential	absolut	0	unbedingt	0.166	unbedingt notwendig	0.10
essential	wesentlichen	0.83	wesentliche	0.909	ist	0
office	Büro	1	Amt	1	Büros	0.8
pollutants	Schadstoffe	1	Schadstoffen	0.916	Emission	0
expertise	Fachwissen	0	Sachverstand	1	Sachkenntnis	1
prescribed	vorgeschrieben	1	vorgeschriebenen	0.875	vorgeschriebene	0.93
means	bedeutet	1	Mittel	1	heißt	0.09
bill	Gesetzentwurf	0.15	Gesetzesentwurf	0.133	Rechnung	1
approach	Ansatz	1	Konzept	0	Vorgehensweise	0
audit	Prüfung	0	Audit	1	Rechnungsprüfung	1

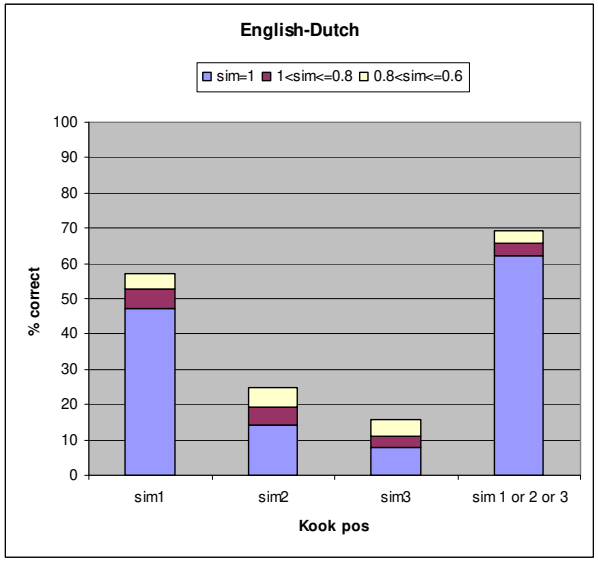
co1-3: top trans-co-occurrences, p1-3: largest prefix match with some dict. entry of “word“.

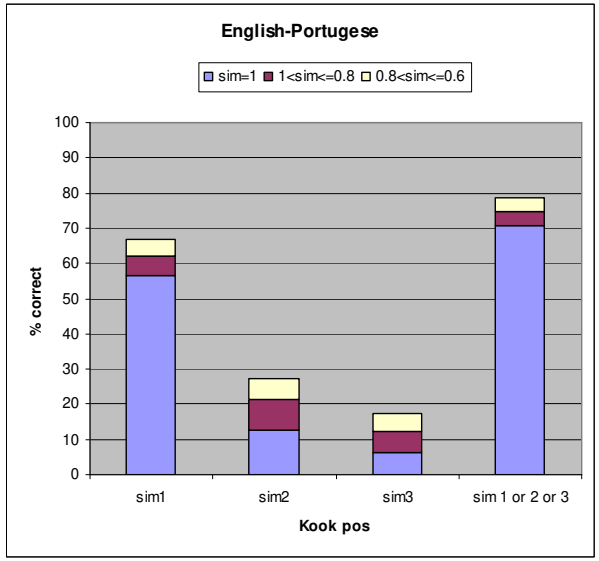
Results for freelang-evaluation

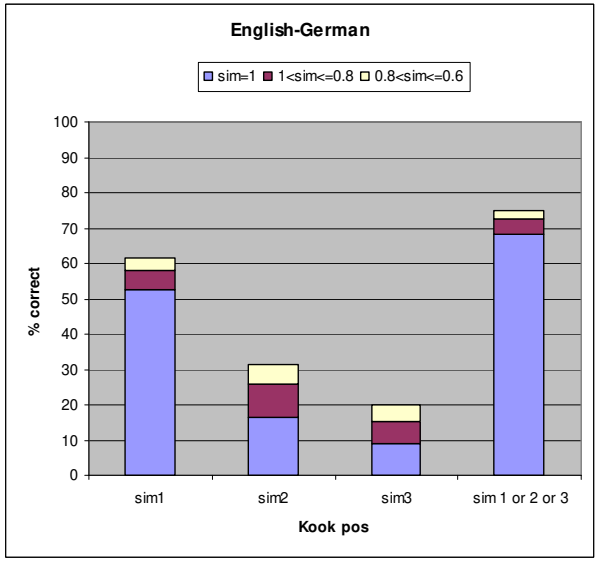








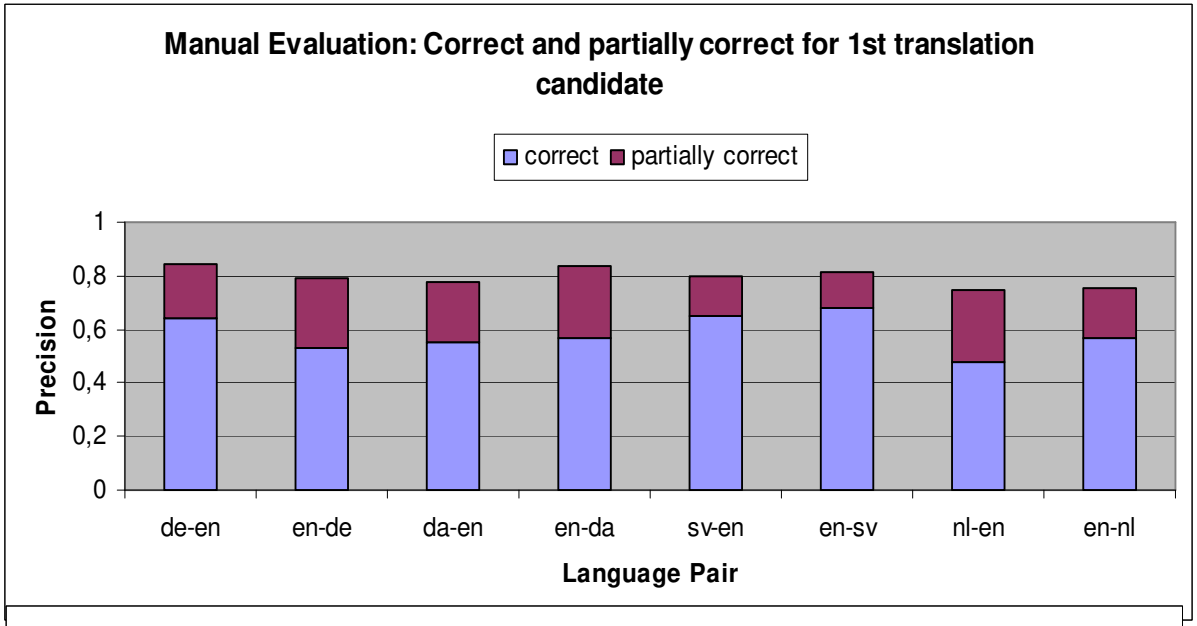


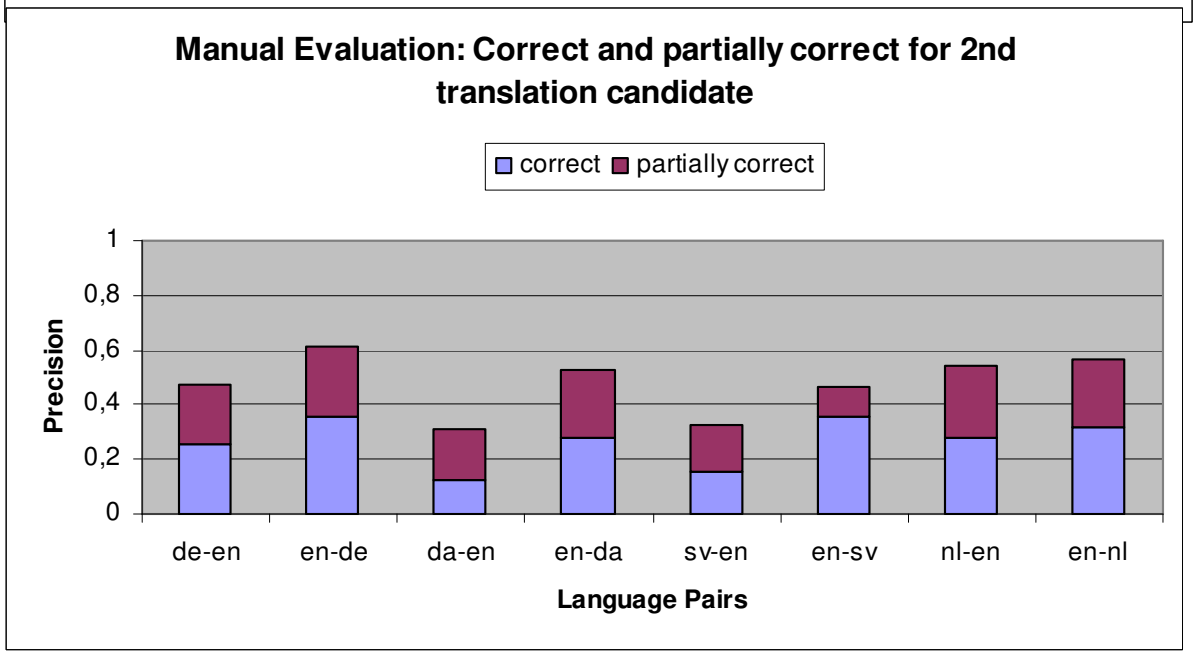


blue: prfx=1, red: 1<prfx<=0.8, yellow: 0.8<prfx<=0.6

Manual Evaluation

on 1000 words random samples



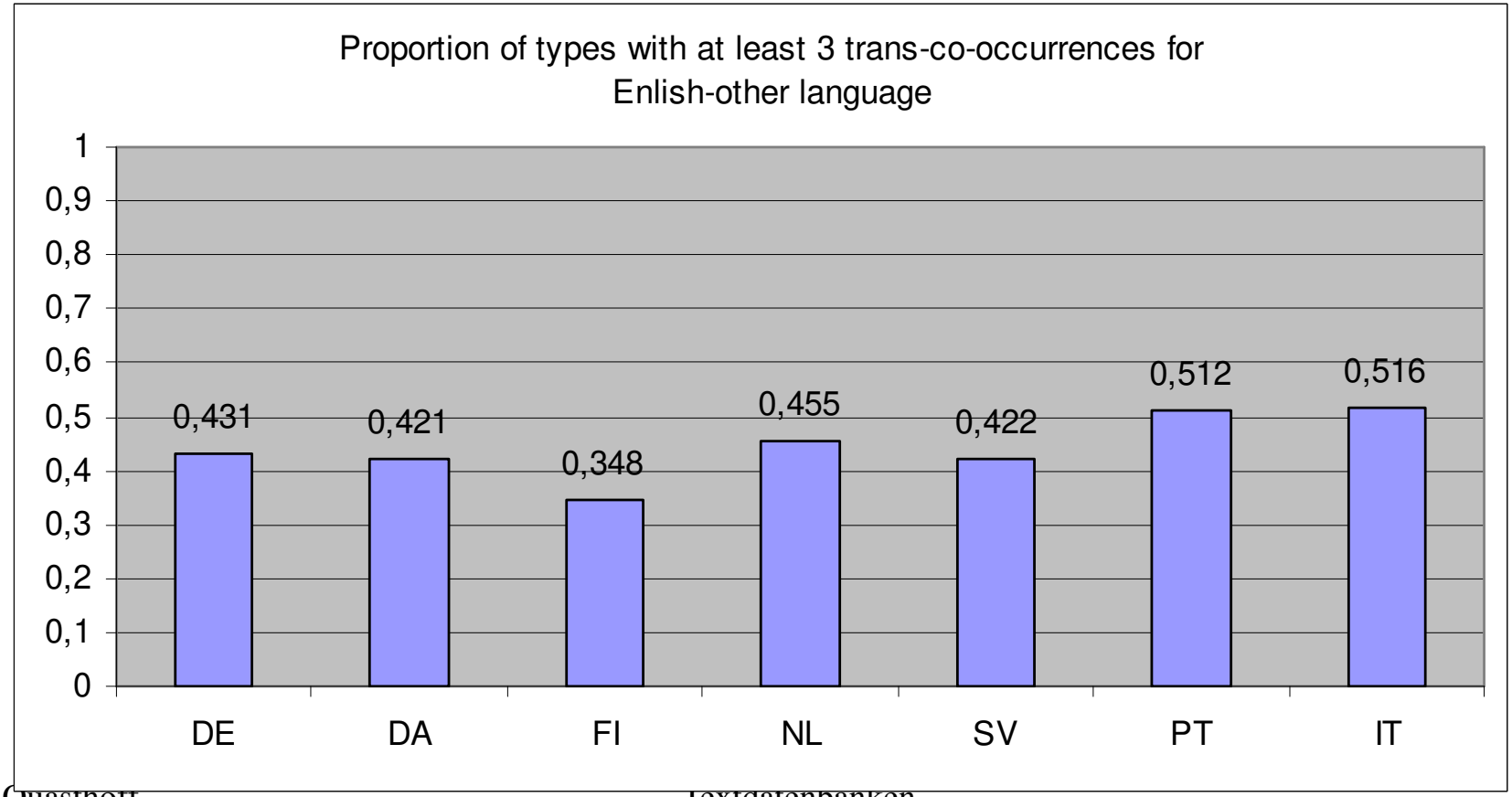


Better results:

- no domain-dependent deficiency of dictionary
- no problems with inflection

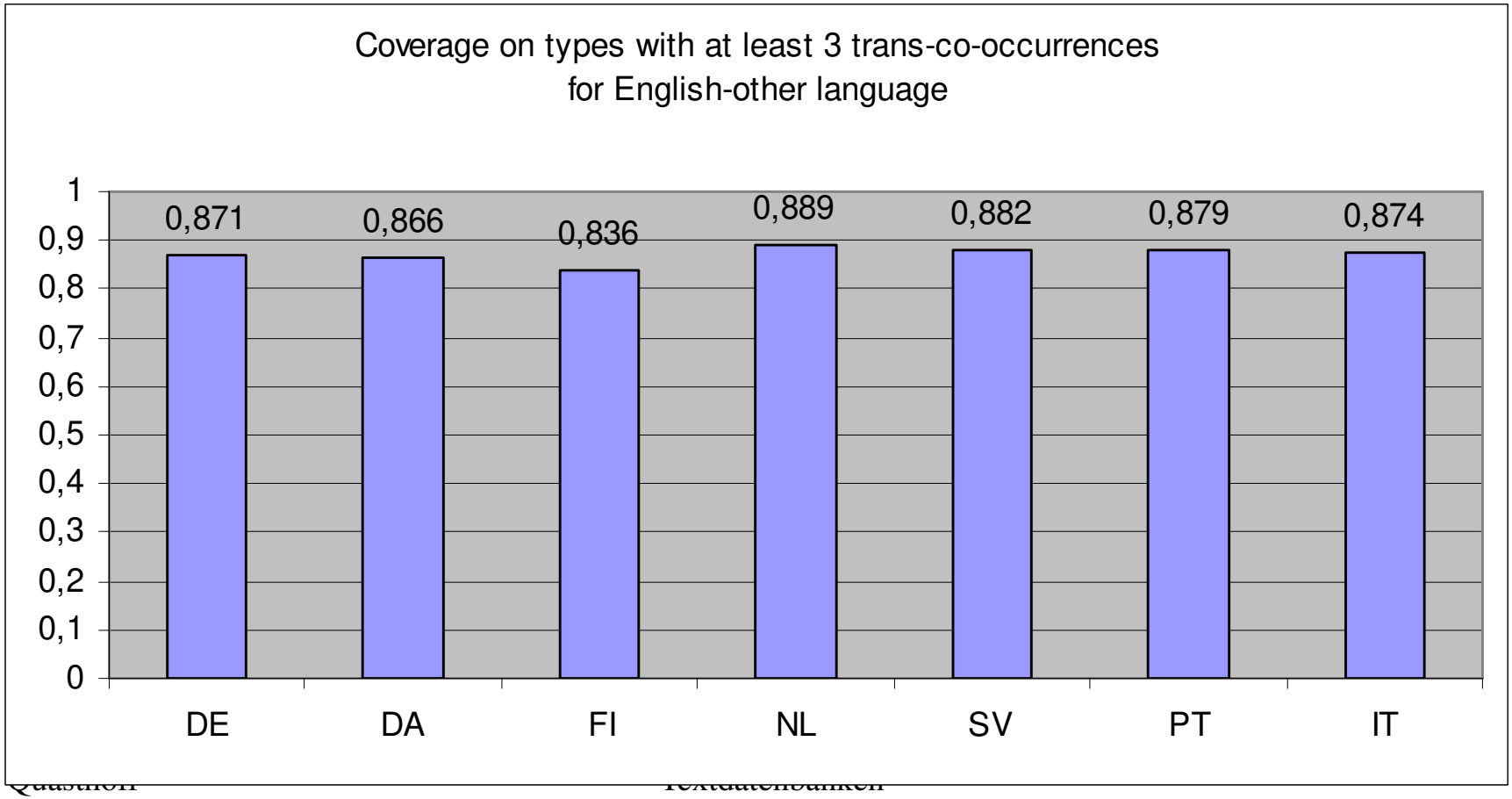
Coverage on types

Proportion of words with at least 3 trans-co-occurrences in types list

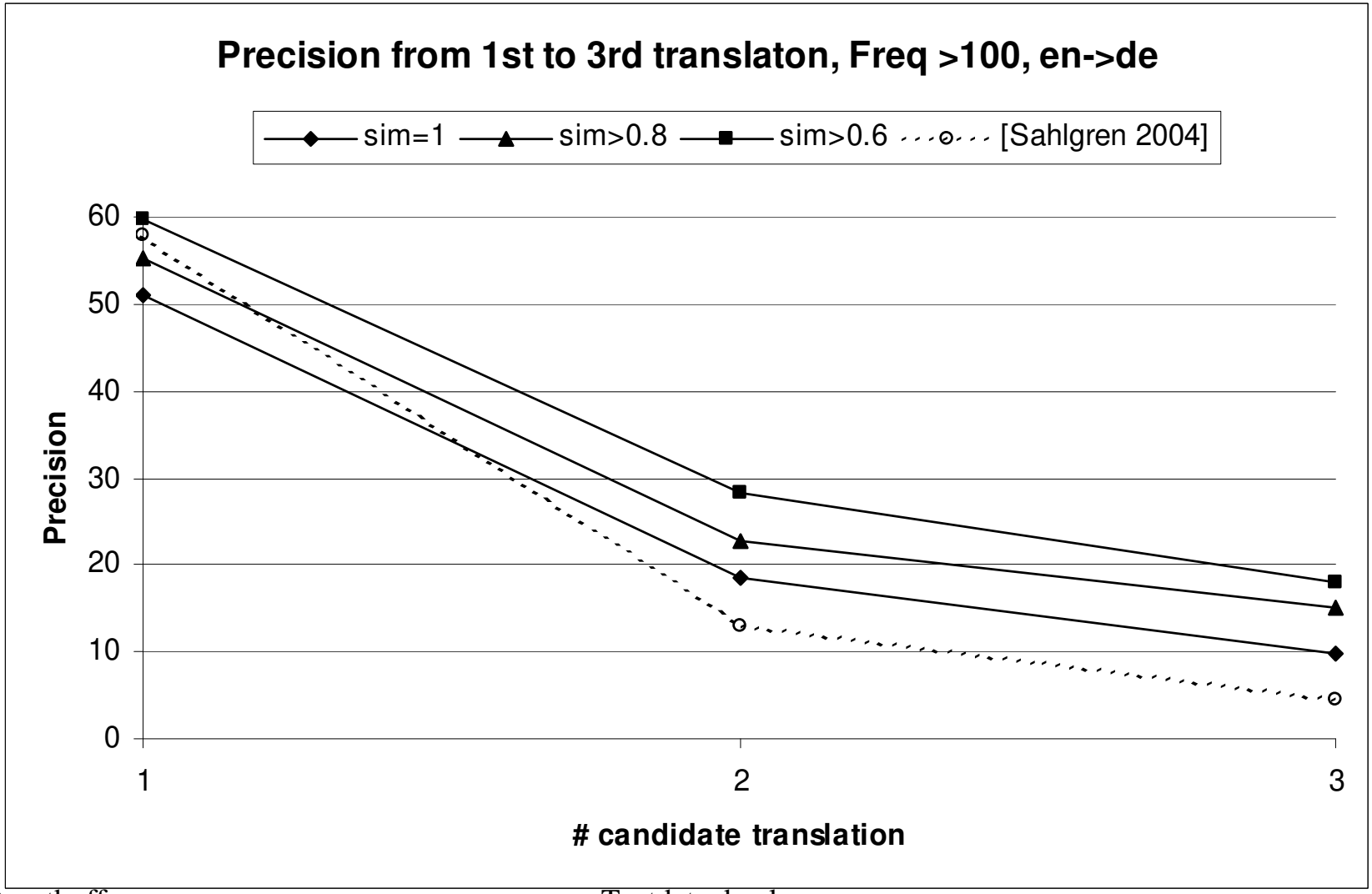


Coverage on tokens

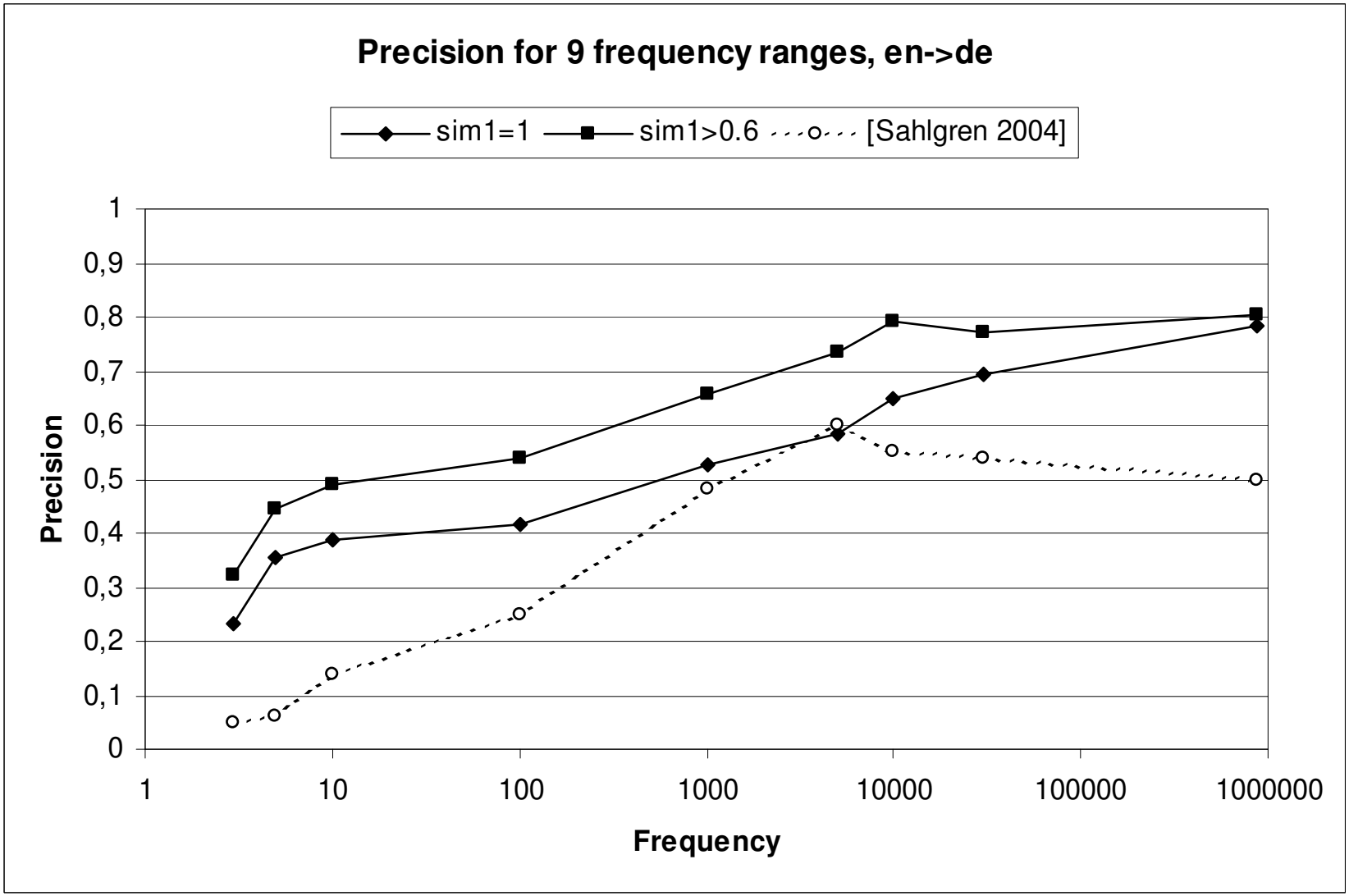
Proportion of tokens having at least 3 trans-co-occurrences in running text.



Comparison with [Sahlgren 2004]



Comparison with [Sahlgren 2004]



Alignment

Given:

- Bilingual sentence pair

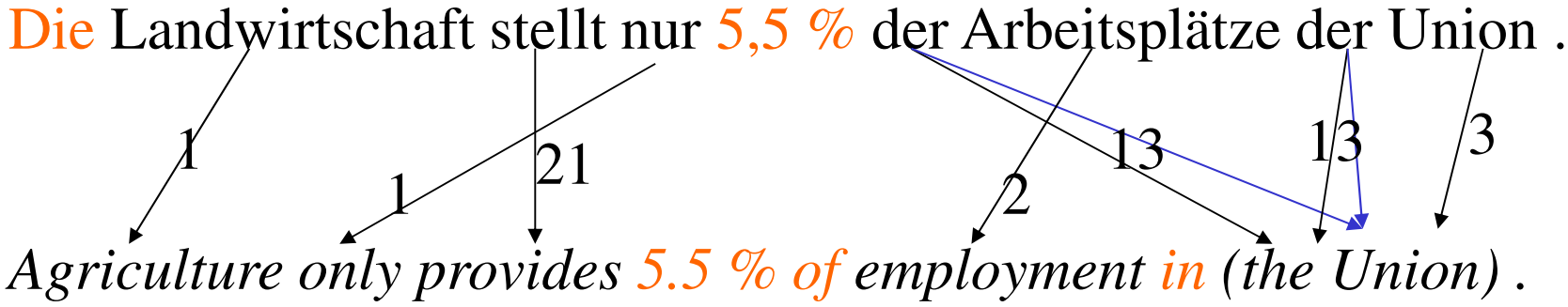
Wanted:

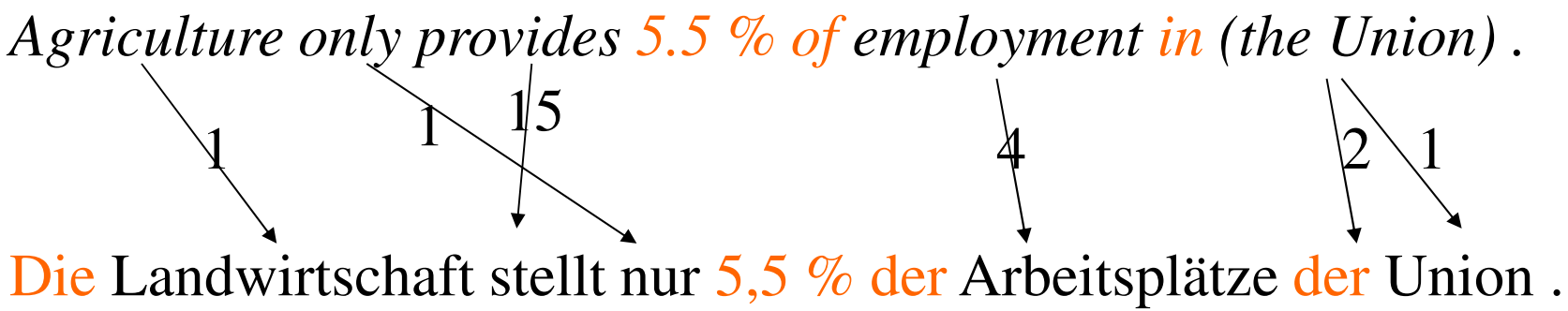
- Which word corresponds with which?

Method:

- Scan sentence 1 word by word and link it to the highest ranked word in the trans-co-coccurrences that can be found in sentence 2.

Alignment: Example 1



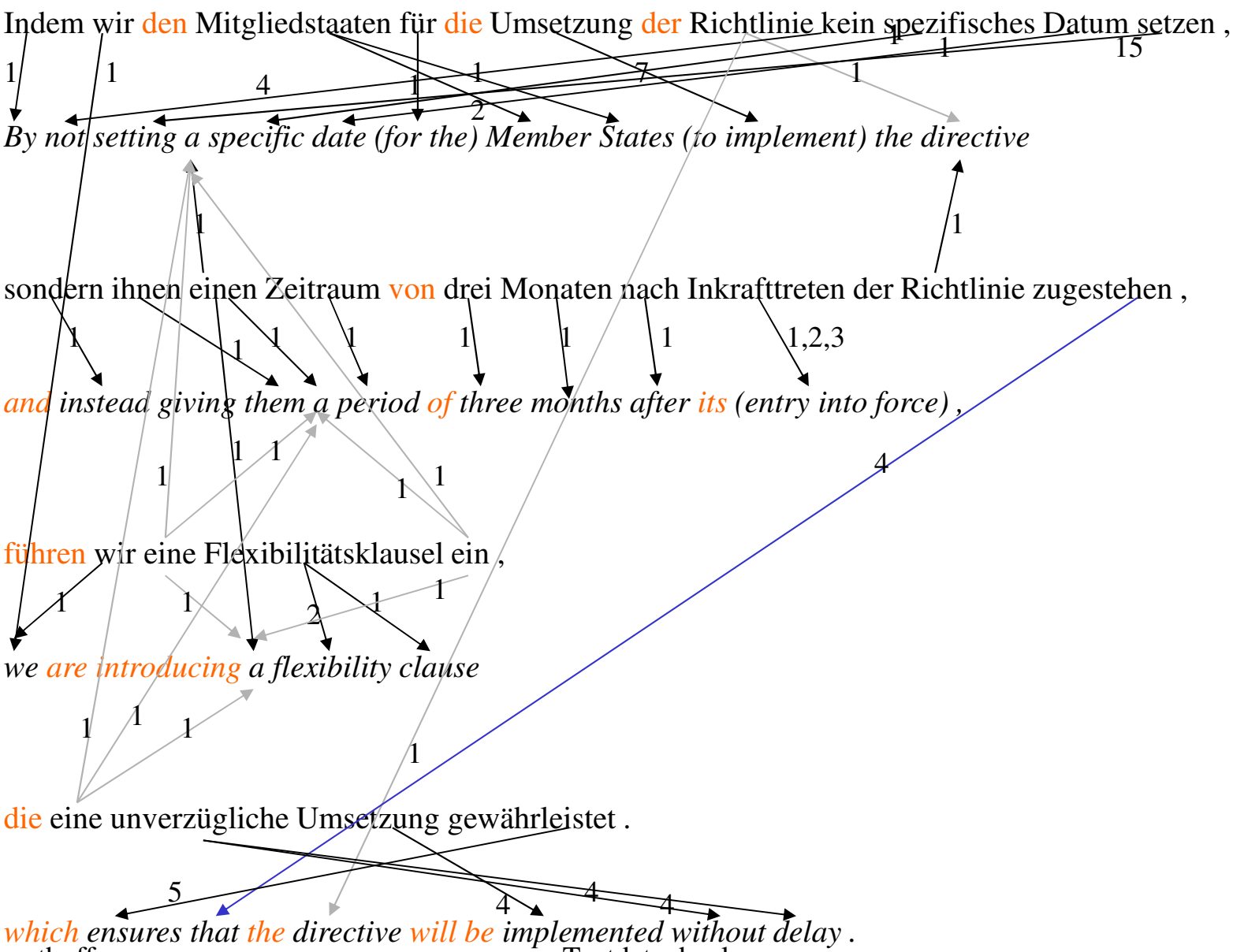


Red Words: No alignment

Blue Arrows: Errors

Arrow Index: rank in trans-co-occurrences

Alignment: Example 2



Grey Arrows: Multiple alignments for frequent words.

Further work

Dictionary acquisition:

- document-level aligned texts
- weakly parallel texts or corpora

Alignment:

- Dealing with cognates
- Symmetric alignment
- Alignment of phrases and multiword units

Wörter für Sachgebiete

Das Problem, typische Wörter für Cluster von Sätzen zu finden, haben wir schon gelöst mit dem Korpus *rubriken (http://wortschatz.uni-leipzig.de/rubrik/).*

Bei den für den Wortschatz ausgewerteten Zeitungstexten war häufig die entsprechende Rubrik (z.B. Politik, Wirtschaft, Auto, Medizin, Hamburg, ...) bekannt. An jeden Satz, für den eine Rubrik bekannt ist, wird der entsprechende Rubrikname angehängt. Danach werden wie gewöhnlich Satzkookkurrenzen ausgerechnet. Die Satzkookkurrenzen der Rubriknamen sind Wörter, die für diese Rubrik typisch sind. In den Kookkurrenzlisten stehen viele hochrelevante Wörter, aber auch Stoppwörter. Hier die Liste für *ssg-Wissenschaft* mit den Stoppwörtern *können* und *etwa*.

Signifikante Kookkurrenzen für "ssg-Wissenschaft": Forscher (8863), Wissenschaftler (3841), Patienten (2949), Zellen (1878), Nature (1548), Studie (1250), Erde (1201), Science (1151), Mediziner (1146), Bd (1140), etwa (1135), Studien (1132), können (1112), Gehirn (1104), Gene (1101), Bakterien (1077), Therapie (952), Tiere (951), Universität (909), ...

Diese Stoppwörter zeichnen sich dadurch aus, dass sie in mehreren Rubriken auftauchen:

Signifikante Kookkurrenzen für "können": ssg-Wissenschaft (1112), ssg-Webwelt (456), ssg-Beruf u. Karriere (185), ssg-Tips & Tricks (174), ssg-WebService (146), ssg-Innovation (96), ssg-Hochschule (75), ssg-Medizin (55), ssg-Berufswelt (43), ssg-Web-Welt Spezial (42), ssg-WebWirtschaft (37), ssg-WebFinanzen (32), ...

Die Wortmengen sind nicht mehr disjunkt, die Bildung von Hierarchien möglich.

term: ssg-Sport
number of occurences: 0
class of frequency: 24 (i.e. *der* has got about 2^{24} the number of occurences than the selected word.)

significant cooccurrences of ssg-Sport:

Trainer (17155), gegen (11004), Spieler (10147), Mannschaft (10141), Tore (8716), Saison (8639), Spiel (8081), m (7415), Bundesliga (7253), WM (6263), Sieg (6165), Fußball (6134), Team (5843), beim (5775), Weltmeister (5333), Finale (5096), Hertha (5035), Leverkusen (4771), Rennen (4508), Dortmund (4483), Fans (4317), Minute (4313), Tor (4269), gewann (4105), Schalke (3956), Minuten (3893), Stürmer (3759), Bayern (3737), Platz (3730), Liga (3700), Halbfinale (3602), Schiedsrichter (3600), Ball (3494), Nationalspieler (3448), Training (3415), DFB (3387), Schumi (3319), Eisbären (3318), Spieltag (3247), Zuschauer (3201), Spiele (3195), Ferrari (3117), Meister (2979), Sydney (2926), Bundestrainer (2918), Daum (2836), Runde (2763), Schumacher (2730), Matthäus (2724), EM (2695), Viertelfinale (2672), Vogts (2642), Kapitän (2635), TeBe (2594), spielen (2591), Partie (2580), Coach (2538), Treffer (2509), Alba (2471), Sekunden (2469), Röber (2426), Punkte (2370), Hoeneß (2349), Nationalmannschaft (2340), Teams (2270), Turnier (2218), Spielen (2187), Klub (2150), Italien (2134), Olympiasieger (2133), Zuschauern (2102), Meisterschaft (2099), Olympia (2076), Atlanta (2060), Manager (2031), km (2022), Torwart (1967), Klubs (1963), Niederlage (1959), Profis (1943)

cooccuring multi words:

Hertha BSC (4077), Champions League (2833), FC Bayern (2591), Michael Schumacher (2497), Bayer Leverkusen (2214), Bayern München (2158), Borussia Dortmund (2114), Gelbe Karten (2106), VfB Stuttgart (2072), Tennis Borussia (2063), Hamburger SV (1743), Steffi Graf (1724), Formel 1 (1643), Lothar Matthäus (1520), Berti Vogts (1513), Werder Bremen (1434), FC Köln (1408), Boris Becker (1339), FC Kaiserslautern (1307), FC Bayern München (1298), Alba Berlin (1289), Hansa Rostock (1269), Real Madrid (1253), Eintracht Frankfurt (1245), Jan Ullrich (1192), FC Barcelona (1161), Franz Beckenbauer (1148), Christoph Daum (1118), Olympischen Spielen (1072), VfL Bochum (1057), Karlsruher SC (1055), FC Union (1050), Tour de France (1049), Otto Rehhagel (1046), Erich Ribbeck (1032), VfB Leipzig (1018), AC Mailand (1005), Mario Basler (959), Mika Häkkinen (919), Berlin Capitals (916), VfL Wolfsburg (905), Jürgen Klinsmann (889), FC Berlin (867), auf der Bank (866), Großen Preis (861), SC Freiburg (860), Oliver Bierhoff (841), Stefan Effenberg (837), MSV Duisburg (830), Rudi Völler (800), SC Berlin (793), Manchester United (740), Ralf Schumacher (729), Olympischen Spiele (724), Erik Zabel (715), Borussia Mönchengladbach (708), Uli Hoeneß (698), Svetislav Pesic (687), Inter Mailand (681), Anke Huber (676), Ottmar Hitzfeld (675), Axel Schulz (674), SC Charlottenburg (672), Nicolas Kiefer (670), Gruppe B (668), Arminia Bielefeld (662), Juventus Turin (660), Heinz-Harald Frentzen (655), Michael Preetz (653), Fortuna Köln (652), Gruppe A (641), FC Nürnberg (640), Matthias Sammer (639), Rote Karte (633), David Coulthard (625), Gelb-Rote Karte (623), Andreas Möller (618), Martin Schmitt (617), Mehmet Scholl (614), Pete Sampras (611)

term: ssg-Leute
number of occurences: 0
class of frequency: 24 (i.e. *der* has got about 2^{24} the number of occurences than the selected word.)

significant cooccurrences of ssg-Leute:

Liebe (7476), Sie (6735), Beruf (6217), Ich (2182), Geld (2108), ich (1882), BILD (1784), Gesundheit (1534), Sex (1468), Privat (1188), Jetzt (1085), mich (1011), Ihre (968), Verona (930), Frau (866), Ihr (822), mir (798), Foto (747), SAT (746), sexy (713), mal (713), Baby (694), Ehefrau (678), Freundin (633), s (626), Ehe (618), Diana (587), Prinzessin (580), ihr (505), l (484), Prinz (474), Lebensjahr (471), Fergie (462), gibt's (455), du (454), Freund (453), Busen (451), Scheidung (424), Lassen (406), schöne (405), meine (405), bin (387), Naddel (381), mein (379), Partner (378), TV-Star (377), Ihnen (374), Michelle (356), Jemand (350), Fotos (349), Madonna (348), Jenny (345), Ehemann (342), hab (333), Ihren (325), Sohn (324), Claudia (318), Romy (317), Mann (312), Mama (312), verheiratet (308), Frauen (308), heiraten (307), Papa (301), Mutter (300), SIE (295), Hochzeit (295), RTL (294), dich (292), Vater (289), Bohlen (283), Mel (279), Juhnke (279), Geri (272), Playboy (267), Dodi (263), Tochter (256), Mallorca (256), Top-Model (254), Männer (251)

cooccuring multi words:

Verona Feldbusch (608), Dieter Bohlen (597), Los Angeles (583), Michael Jackson (321), Harald Juhnke (305), Jenny Elvers (298), Prinz Charles (280), Prinzessin Diana (275), Claudia Schiffer (263), Pamela Anderson (256), Mick Jagger (242), Thomas Gottschalk (228), Heiner Lauterbach (217), Spice Girls (216), im Bett (201), z. B. (185), Big Brother (165), Harald Schmidt Show (157), Birgit Schrowange (156), Jerry Hall (153), Modern Talking (152), Ernst August (142), Tommy Lee (141), Rex Gildo (139), Jürgen Drews (138), Guildo Horn (136), Günter Strack (133), Udo Jürgens (130), Drafi Deutscher (130), am Strand (127), Tic Tac Toe (126), Naomi Campbell (125), Götz George (124), Spice Girl (123), Demi Moore (121), Matthias Reim (120), Susan Stahnke (116), Stefan Raab (113), Michael Douglas (112), Cindy Crawford (112), Sharon Stone (109), Monica Lewinsky (108), Jennifer Lopez (107), seine Frau (106), Harald Schmidt (106), Nicole Kidman (105), Whitney Houston (104), Tom Cruise (99), Hans Meiser (97), Elton John (97), Dagmar Berghoff (97), Brad Pitt (97), Liz Taylor (96), Britney Spears (96), ins Bett (95), Til Schweiger (95), Rod Stewart (94), Kelly Family (93), Kate Moss (93), Eva Herman (92), Bruce Willis (90), 7 Tage - 7 Köpfe (90), ein Paar (87), Rudi Carrell (86), Roland Kaiser (85), Margarethe Schreinemakers (85), Sylvester Stallone (83), Sonja Kirchberger (83), Otto Waalkes (83), Christian Anders (83), Roy Black (82), Jeanette Biedermann (82), Puff Daddy (81), Jack Nicholson (81), Ernst August von Hannover (80), ich habe (79), Zeit für (79), Veronica Ferres (79), Patrick Lindner (79), Kai Pflaume (79)

term: ssg-Medizin
number of occurences: 0
class of frequency: 24 (i.e. *der* has got about 2^{24} the number of occurences than the selected word.)

significant cooccurrences of ssg-Medizin:

Patienten (591), Forscher (508), Stammzellen (326), Zellen (288), Gene (198), Wissenschaftler (193), Behandlung (187), Bakterien (183), Therapie (172), Studie (165), Medikamente (161), dass (157), Gehirn (148), Viren (146), Gewebe (141), Studien (132), Gen (122), Körper (119), Mediziner (114), Professor (111), Mäusen (110), Medikament (108), Menschen (96), Protein (91), Frauen (89), Antibiotika (87), Babys (81), Immunsystem (80), Alzheimer (79), genetische (77), Nervenzellen (76), Infektion (76), Ärzte (75), Patient (74), Gigliola (74), klinischen (73), Proteine (73), bei (72), Impfstoff (72), Forschung (71), Virus (70), Krankheiten (69), Nebenwirkungen (68), Zelle (66), Therapien (66), Erbgut (65), ik (64), embryonalen (64), University (62), Krankheit (62), menschlichen (61), Universität (61), Eizellen (60), Ergebnisse (58), genannte (56), Tumorzellen (56), Substanzen (56), Schmerzen (56), können (55), entwickeln (55), Schmerz (55), Methode (54), gesunden (53), Wirkstoff (53), Salmonellen (53), Operation (53), Krebs (53), Symptome (52), Enzym (52), bestimmte (51), Mutterleib (51), Heilung (51), Organismus (50), Mäuse (50), Hirntumoren (50), diese (49), Resistenzen (49), Diabetes (49), Aids (49), Wirkstoffe (48)

cooccuring multi words:

so genannten (60), im Blut (47), San Francisco (28), University of California (23), John Collinge (20), behandelt werden (19), British Medical Journal (19), an der Universität (17), nach der Geburt (16), mit Hilfe (16), New Scientist (16), School of Medicine (14), New Orleans (14), in Zukunft (13), auf der Oberfläche (13), zum Beispiel (12), auf Grund (12), Medical School (12), Berthold Schneider (12), im Alter zwischen (11), University College (11), Proceedings of the National Academy of Sciences (11), Chorea Huntington (11), eine Rolle spielen (10), ein Patient (10), auf den Markt kommen (10), Johns Hopkins Universität (10), Fort Collins (10), Dietrich Grönemeyer (10), Die Ärzte (10), vom Affen (9), University of Southern California (9), National Institutes of Health (9), Medical Center (9), zum Schutz vor (8), zu Grunde (8), unter anderem (8), unter Leitung von (8), eines Tages (8), University of Texas (8), University of Michigan (8), Thomas von Aquin (8), The Lancet (8), La Jolla (8), Jochen Senges (8), General Hospital (8), Frank Goebel (8), zu Lande (7), viel versprechend (7), in vielen Fällen (7), im Übermaß (7), entdeckt werden (7), ein Segen sein (7), bindet an (7), Lennart Nilsson (7), Kary Mullis (7), Gerhard Müller (7), zur Vorbeugung (6), wie zum Beispiel (6), von Tieren (6), unter die Haut (6), ohne dass (6), noch nicht (6), nicht bewusst (6), neues Verfahren (6), neue Methode (6), mit Pausen (6), künstliche Befruchtung (6), in der Lage (6), etwa fünf (6), ein Schlaganfall (6), durchgeführt werden (6), bisher nicht (6), bis zu (6), außer Gefecht setzen (6), auf diese Weise (6), University of Maryland (6), Schmerzen lindern (6), Ohio State University (6), Norbert Lossau (6)

Übersetzen mit 800 Bibeln

- Jedes Wort wird beschrieben durch einen Vektor der Dimension (rund) 8.000. Dieser beschreibt, in welchen Versen ein Wort auftritt.
- Annahme: Übersetzungspaare haben eine fast identische Verteilung. Bei Mehrdeutigkeiten bleibt die Verteilung hoffentlich noch ähnlich.
- Ähnlichkeitsmaß: Skalarprodukt der normierten (!) Vektoren.

Test: DEU – DEU

use deu-erben_bible_2012

```
select w.word, w2.word, round(100*count(distinct iso.s_id)/sqrt(w.freq)/sqrt(w2.freq),2) as quot from words w, inv_w iw, inv_so iso, sources so, `deu-neue_bible_2012`.words w2, `deu-neue_bible_2012`.inv_w iw2, `deu-neue_bible_2012`.inv_so iso2, `deu-neue_bible_2012`.sources so2  where w.w_id=iw.w_id and iw.s_id=iso.s_id and iso.so_id= so.so_id and so.source=so2.source and w2.w_id=iw2.w_id and iw2.s_id=iso2.s_id and iso2.so_id= so2.so_id and w.word="Mensch" group by w2.word order by quot desc limit 10;
```

Mensch	Mensch	39.30
Mensch	Mann	23.35
Mensch	verpachtete	15.02
Mensch	ein	13.25
Mensch	strenger	13.01
Mensch	Sage	12.26
Mensch	verirrt	12.26
Mensch	Wachtturm	12.26
Mensch	forderst	12.26
Mensch	Gegenwert	12.26

Mann	Mann	37.57
Mann	Ehemann	17.15
Mann	Frau	16.52
Mann	teilgenommen	14.85
Mann	Stirbt	14.00
Mann	gebildeter	14.00
Mann	5000	14.00
Mann	namens	13.10
Mann	gerechter	12.13
Mann	oberste	11.43

Zeugnis	Zeugnis	38.48
Zeugnis	bezeugen	23.26
Zeugnis	bezeugt	18.53
Zeugnis	Falschaussagen	18.16
Zeugnis	zeig	18.16
Zeugnis	bestätigt	15.25
Zeugnis	Zeugenaussage	14.82
Zeugnis	größeres	14.82
Zeugnis	stimmten	14.82
Zeugnis	angewiesen	14.82

alsbald	Augenblick	25.73
alsbald	Sofort	23.72
alsbald	verschwand	22.36
alsbald	sofort	22.36
alsbald	selben	18.61
alsbald	Gleich	16.77
alsbald	Aussatz	16.77
alsbald	Im	16.01
alsbald	Sobald	15.81
alsbald	nachher	15.81

1:2 Komposita

Mutterleib	mother's	65.47
Mutterleib	womb	46.29
Mutterleib	Beautiful	40.82
Mutterleib	separated	23.57
Mutterleib	alms	19.25
Mutterleib	lame	16.01
Mutterleib	daily	14.91
Mutterleib	pleasure	14.00
Mutterleib	carried	13.25
Mutterleib	strong	11.32

womb	Mutterleib	46.29
womb	hüpfte	37.80
womb	Almosen	26.73
womb	gesegnetste	26.73
womb	herbeigetragen	26.73
womb	Elisabets	26.73
womb	vernahm	26.73
womb	berauschenden	26.73
womb	verzichten	26.73
womb	Getränke	26.73

mother's	Mutterleib	65.47
mother's	Schöne	37.80
mother's	herbeigetragen	37.80
mother's	anrühren	37.80
mother's	gelähmt	37.80
mother's	verzichten	37.80
mother's	Klopas	37.80
mother's	Getränke	37.80
mother's	berauschenden	37.80
mother's	Almosen	37.80

Wortstellung: NB-Kookkurrenzen

Beispiel: „ewiges Leben“ (sig=312)

ewiges	eternal	58.93
ewiges	life	30.81
ewiges	believes	21.21
ewiges	disobeys	20.00
ewiges	incorruptibility	14.14
ewiges	mid	14.14
ewiges	snatch	14.14
ewiges	well-doing	14.14

Leben	life	64.45
Leben	eternal	44.48
Leben	Good	15.94
Leben	death	15.22
Leben	lifetime	13.02
Leben	causes	11.64
Leben	believes	11.39
Leben	wasted	10.63

```
use eng-web_bible_2012
select c.*, w1.word, w2.word from co_n c,
words w1, words w2 where w1.w_id=w1_id
and w2.w_id=w2_id and w1.word="eternal"
and w2.word="life" limit 30;
```

w1_id	w2_id	freq	sig	word	word
425	224	44	520.66	eternal	life

Übersetzen über mehrere Zwischensprachen

- Um ein Wort von Sprache A nach Sprache C zu übersetzen, wird es zunächst in viele Sprachen B1, B2, … übersetzt und diese Übersetzungen dann nach Sprache C.
- Hoffnung: Die korrekte Übersetzung von A tritt am häufigsten in der Liste in Sprache C auf.
- Das ist sinnvoll, falls es keine „direkte“ Übersetzung von A nach C gibt.

Start: SWE: Jesus (1)

```
select  left(t2.DB2,3) as lang2,t2.db2, t2.word2, sum(t1.sim*t2.sim) as wert from trans t1, trans t2 where t1.DB1 like "swe%" and
t1.DB2 not like "deu%" and t1.DB2 not like "eng%" and t1.DB2=t2.DB1 and t1.w2_id=t2.w1_id and t1.word1="Jesus" group by t2.DB2,
t2.word2 order by wert desc limit 100;
```

lang2	db2	word2	wert
swe	swe-sfb_bible_2012	Jesus	25.6170351755285
por	por-001_bible_2012	Jesus	25.1083889329419
ron	ron-011_bible_2012	Isus	24.8066131365056
spa	spa-bda_bible_2012	Jesús	24.4006532003255
hau	hau-001_bible_2012	Yesu	24.1914100617731
ceb	ceb-cbv_bible_2012	Jesus	24.105260083189
ita	ita-dio_bible_2012	Gesù	24.0279981299896
dan	dan-1871_bible_2012	Jesus	23.9889311481359
afr	afr-1983_bible_2012	Jesus	23.9559769380074
eng	eng-bishops-wbt_bible_2012	Iesus	23.9407731443067
vie	vie-1934_bible_2012	Jêsus	23.9155320951097
yor	yor-yce_bible_2012	Jesu	23.8984500431895
fra	fra-dby_bible_2012	Jésus	23.8660601348305
eng	eng-coverdale-wbt_bible_2012	Iesus	23.6889201025236
deu	deu-neue_bible_2012	Jesus	23.6191960331895
ind	ind-bar_bible_2012	Yesus	23.540775104022
eng	eng-bbe_bible_2012	Jesus	23.380954046782
nob	nob-001_bible_2012	Jesus	23.2947209268072
pol	pol-001_bible_2012	Jezus	23.2286652002847

Start: SWE: Jesus (2)

lang2	db2	word2	wert
deu	deu-luther1912_bible_2012	Jesus	23.2279590576205
deu	deu-albrecht_bible_2012	Jesus	23.2047129783626
rus	rus-ibs_bible_2012	Иса	23.1408491544793
hye	hye-w-1853_bible_2012	Yisows	23.1066229932268
deu	deu-fb2004_osis_bible_2012	Jesus	23.0998401497638
ces	ces-bkr_bible_2012	Ježíš	23.0173270529728
pes	pes-tpv_bible_2012	22.9254990224371	عیسی
cat	cat-ev_bible_2012	Jesús	22.9179351070845
som	som-sim_bible_2012	Ciise	22.846520099293
deu	deu-luther1545_bible_2012	Jesus	22.8225669796185
lit	lit-001_bible_2012	Jėzus	22.8196559212091
bul	bul-b40_bible_2012	Исус	22.8140480268974
hrv	hrv-001_bible_2012	Isus	22.7779630134995
deu	deu-abraham_meister_bible_2012	Jesus	22.7718340948436
deu	deu-elb1871_bible_2012	Jesus	22.6686111471071
aln	aln-1990_bible_2012	Jezusi	22.6506139773355
als	als-abv_bible_2012	Jezusi	22.6506139773355
aze	aze-bsa_bible_2012	İsa	22.6222660841968
tur	tur-001_bible_2012	İsa	22.476089047955
tgl	tgl-mbb_bible_2012	Jesus	22.4123710198023
kaz	kaz-kaz_bible_2012	Иса	22.4035260961583

Quasiparalleler Text

Definition:

Zwei Texte in verschiedenen Sprachen heißen quasiparallel, wenn ihre Inhalte (nahezu) übereinstimmen, aber ein Text nicht notwendig als Übersetzung des anderen angesehen werden kann.

Mögliche Unterschiede liegen in

- der Reihenfolge im Text
- kleinen inhaltlichen Abweichungen (Ergänzungen, Auslassungen, ...)

Frage: Lassen sich auch solche quasiparallelen Texte zur Wörterbucherzeugung verwenden?

Hoffnung: Es gibt viel quasiparallelen Text, z.B. Agenturmeldungen, verschiedene Wikipedias

Beispiele

http://www.cumberlink.com/articles/2006/11/16/ap/business/d8le5m782.txt

BENTONVILLE, Ark. - Wal-Mart Stores Inc. Said Thursday it is expanding its $4 generic prescription program to 11 additional states and adding 17 more prescriptions to the program.The world's largest retailer added 502 stores to those offering the discounted medications. The new states added Thursday were Idaho, Kentucky, Maine, Massachusetts, Nebraska, Oklahoma, Rhode Island, South Carolina, Utah, Washington and West Virginia. In all, Wal-Mart is offering the program in 3,009 stores in 38 states.The $4 price is for up to a 30-day supply of the drugs, which will now number 331. That number counts some drugs more than once if they are sold in a variety of dosages or solid and liquid forms.The company began the program in September, offering the low-cost drugs in Florida and had plans to expand the offering in January. But the company said it moved up its timetable.Wal-Mart launched the program in what it called an effort to save working Americans money on health care. Critics said it was a stunt to draw in business and a grab for a bigger share of the drug business.

http://www.stock-world.de/news/article.m?news_id=2178274

Wal-Mart erweitert $4-Generika-Programm Börsennachrichten - Börse aktuell

Der weltgrößte Einzelhändler Wal-Mart Stores Inc. hat sein Programm des Verkaufs rezeptpflichtiger Generika für 4 Dollar auf 11 weitere US-Bundesstaaten ausgedehnt. So sind 502 Niederlassungen in Idaho, Kentucky, Maine, Massachusetts, Nebraska, Oklahoma, Rhode Island, South Carolina, Utah, Washington und West Virginia vorgesehen. Dem in 38 Bundesstaaten in 3009 Wal-Mart-Apotheken angebotenen Programm werden 17 Medikamente hinzugefügt. Es umfasst nun 331 Präparate. Die Verschreibungsliste beinhalte 14 der 20 am häufigsten verordneten Arzneiverschreibungen in den USA. Der Startschuss für das Programm fiel durch Wal-Mart im September in Florida.

http://news.yahoo.com/s/afp/20061116/en_afp/afpentertainmentireland_061116124722

LONDON (AFP) - "Lord of the Dance" Michael Flatley is seriously ill in a London hospital, a publicist for the Irish dancer has said. All 20 dates for his sold out Celtic Tiger tour of Europe, due to start in a few days, have been scrapped and the 48-year-old's wife is at his bedside. His publicist said he had been in hospital for some time, while the priest who married the couple said he had been there for two weeks. There are no further details at the moment, but to say he has been in for several days," the publicist said. A statement on Flatley's website read: "Celtic Tiger Touring Inc. has cancelled all European dates for Celtic Tiger starring Michael Flatley due to serious illness.

Born in Chicago to Irish immigrant parents, Flatley shot to fame in the "Riverdance" Irish step dancing spectacular, which took off after being first performed during the 1994 Eurovision Song Contest interval. He created his own hit shows "Lord of the Dance", "Feet of Flames" and "Celtic Tiger". The dancer wed his pregnant second wife Niamh, 32, just four weeks ago. Father Aidan Troy, who married the couple, said he spoke to Flatley in hospital two weeks ago. I spoke to him about two weeks ago and he was in hospital in London," the close friend said.

The Sun newspaper quoted an unnamed friend of Flatley as saying: ...

http://www.westfaelischer-anzeiger.de/afp/storydetail.php?rubr=unterhaltung&rub=spezial--unterhaltung&urb=061116114035.crvqkimn

Der "Lord of the Dance"-Star Michael Flatley liegt schwer krank in einer Londoner Klinik. Sämtliche 20 Termine seiner "Celtic Tiger"-Tournee durch Europa, die in wenigen Tagen beginnen sollte, seien abgesagt, sagte ein Sprecher des Tänzers. Flatleys Frau wache an seinem Krankenbett. Der Tänzer sei schon längere Zeit im Krankenhaus. Die "Sun" zitierte einen Freund des Stars mit den Worten, Flatley sei "sehr schwach". Sein ganzer Körper sei von einer Infektion betroffen, deren genaue Hintergründe noch unklar seien.

Der in Chicago als Sohn irischer Einwanderer geborene Flatley war mit der irischen "Riverdance"-Steptanz-Truppe bekannt geworden. Später stellte er seine eigenen Shows auf die Beine. Erst vor vier Monaten heiratete er seine zweite Frau Niamh, das Paar erwartet ein Kind.

http://english.aljazeera.net/NR/exeres/6D84E985-DAD6-4442-BD67-DCA33D2BD1AB.htm

Segolene Royal is hoping to win the nomination to lead France's socialists into next year's presidential election as party members vote to choose a candidate.

Royal, 53, has enjoyed a large opinion poll lead over her more experienced party rivals Dominique Strauss-Kahn and Laurent Fabius in recent months.

Outright victory would underline her opinion poll status as the only socialist capable of beating the right's Nicholas Sarkozy next year.

Despite a long career in French politics, her image as a fresh face, strong on traditional values and ready to listen to citizens' concerns has played well with a public tired of leaders seen as more elitist.

Strauss-Kahn, 57, a former finance minister running as a social democrat and Fabius, 60, a prime minister in the 1980s now positioned on the party's left wing, have trailed behind her.

A poll in the weekly Le Point on Thursday put Royal and Sarkozy level if they face each other in the decisive round of voting next year.

The outcome of Thursday's poll is unpredictable because opinion polls have counted "socialist sympathisers" rather than the 219,000 actual party members who will vote.

http://www.frankenpost.de/nachrichten/brennpunkte/resyart.phtm?id=1049159

Segolene Royal gilt bei der Kandidatenkür der Sozialisten als favorisiert.

Paris (dpa) - Die populäre französische Regionalpolitikerin Ségolène Royal (53) ist am Donnerstag als große Favoritin in die Wahl des sozialistischen Präsidentschaftskandidaten für 2007 gegangen.

Gegen sie treten zwei Mitbewerber an: der ehemalige Premierminister Laurent Fabius und Ex-Wirtschafts- und Finanzminister Dominique Strauss-Kahn. Die knapp 219 000 Mitglieder der Oppositionspartei sind aufgerufen, von 16.00 bis 22.00 Uhr in 4000 Wahlbüros unter den drei Bewerbern ihre Wahl zu treffen.

Eine am Donnerstag veröffentlichte Umfrage des Ipsos-Instituts unter Sympathisanten der Sozialistischen Partei (PS) ergab mit 66 Prozent für Royal eine um vier Prozentpunkte gestiegene Zustimmung zu der früheren Umwelt- und Familienministerin. Ihre beiden Mitbewerber um die Kandidatur fielen nach der Umfrage entsprechend zurück: Fabius kam nur noch auf 10 Prozent, Strauss-Kahn auf 24 Prozent.

Wie werden die Texte verbunden?

Ziemlich sicher durch

- Eigennamen (bei gleicher Schreibweise)
- Zahlen, Maßeinheiten

Weniger sicher durch

- Cognates

Wir bezeichnen solche Wörter als *Anker*.

Algorithmus zum Finden von Übersetzungspaaren

- Rechts oder links stehende Wörter neben gleichen Ankern sind Kandidaten für Übersetzungspaare.
- Wiederholtes Auftreten in solcher Position innerhalb eines Dokuments erhöht die Sicherheit.

- Verifikation: Überprüfen in anderen Paaren von Dokumenten.

Übersetzungspaare aus nicht-parallelem Text

Gegeben: Vergleichbare Korpora (z.B. Zeitungstext aus dem gleichen Zeitraum) ohne irgendeine Zuordnung der Texte untereinander.

Trotzdem lassen sich Kookkurrenzen hoffentlich oft wörtlich übersetzen.

Beispiel: **EPO**

Unter den stärksten Kookkurrenzen findet man die Cognates

- Erythropoietin, Hämatokritwert, Hormon (de),
- Erütropoietiini, Hematokriidiprotsent, hormooni (ee),
- ERYTHROPOIETIN, hormóninu (is)

Achtung, schwieriger zu verifizieren! (Warum?)