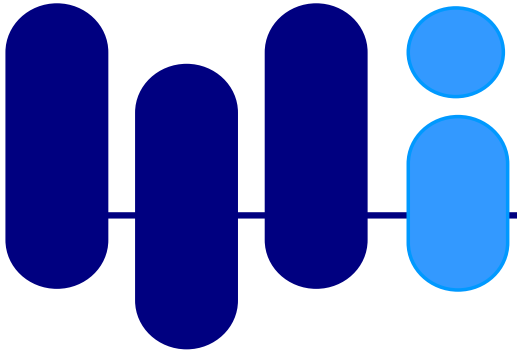
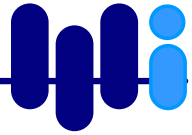


Collocations, Language Change, and Media Analysis

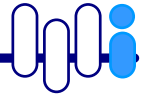
Warsaw Linguistics Workshop, April 2006

Christian Wolff

Regensburg University, Media Computing

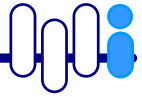


Overview



- Motivation
- Analysis and processing of day-based corpora
- Visualization of Results
- Analyzing change in collocation sets

Acknowledgements



- This presentation is based on work of
  - The natural language processing group at Leipzig University (Prof. Heyer, Prof. Quasthoff et al.)
  - Ulrike Mendel (M.A. student, Regensburg University)
  - some thoughts of my own

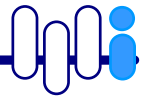
"Wortschatz" Project – Leipzig University

- corpus collection since 1995
- monolingual text corpora: de, en, fr, nl, ...
- sentence-based approach
- basic statistics for inflected as well as non-inflected forms

Basic data of the collections:

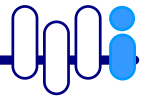
	German	***English***	***Dutch***	***French***
word tokens	300 Mill.	250 Mill.	22 Mill.	15 Mill.
sentences	13,4 Mill.	13 Mill.	1,5 Mill.	860.000
word types	6 Mill.	1,2 Mill.	600.000	230.000

Timeline



- 1995-2000:
  - collection of reference corpus
  - analysis routines and technical infrastructure
- 2000-2003:
  - day-based corpus collection
  - analysis of “words of the day”
- 2004/2005
  - normalized corpora of equal size for individual days / years
  - in different languages
- 2005ff
  - analysis of time-based change in collocation sets

Data Sources

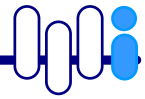


- general newspaper text, e. g.
  - major German newspapers
  - English newspaper text from the TREC / TIPSTER collections)
- electronic dictionaries
  - general purpose language dictionaries, esp. for grammatcial categories, pragmatics information, classification
  - dictionaries for special domains (e. g. electrical engineering, medicine)
- electronic books and journals (CD-ROM-based)
- **web resources: agent-controlled acquisition from "reliable" sources**

Information Categories (German Corpus)

Category	Number of Entries
overall size	~ 300 Mio. words
word list (inflected forms)	~ 7 Mio. different word forms
example sentences	~ 25 Mio.
grammatical information	~ 3 Mio.
morphological information	~ 3 Mio.
descriptions	~ 150.000
subject categories	~ 1,5 Mio.
semantic relations	~ 500.000
pragmatics (e. g. usage)	~ 35.000
collocations (sentence level)	~ 3,5 Mio.
collocations (immediate left and right neighbours)	~ 1,5 Mio.

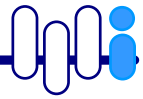
Data Analysis



- pre-processing (conversion)
- sentence segmentation
- token recognition (inflected forms)
- calculation of typical collocations
- quality optimisation by checking of data against high quality sources and through (little)manual checking

- simple and portable suite of tools

Entry Example: *Weltanschauung* (światopogląd)



Word (*word number*: 95400): Weltanschauung

Frequency class: 14 (Absolute count: 387)

Subject Area: General, Chemistry, Natural Science, Science, Culture, Education, Learning, Chemie -> Naturwissenschaft -> Wissenschaft -> Kultur Erziehung Bildung Wissenschaft)

Morphology:welt|an|schau|ung (=welt+an=schau%ung)

Grammatical Information:

Part of Speech: Noun

Gender: Feminine

Inflection: die Weltanschauung, der Weltanschauung, [...] (*inflection class fb*)

Relations to other Entries:

Synonyms: Anschauungsweise, Betrachtungsweise, Denkweise

Compare To: Fatalismus, Idealismus, Ideologie, Kommunismus, Nihilismus, Optimismus, Pazifismus, Realismus

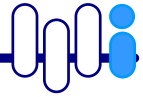
Synonym of: Anschauungsweise, Denkart, Denkungsweise, Denkweise, Einstellung, Ideologie, Lebensanschauung, Meinung, Mentalität, Philosophie, Sinnesart, Standpunkt, Urteil, Weltbild

Examples:

Auch die Schulmedizin beinhaltet schließlich eine Weltanschauung - eben die rein naturwissenschaftliche. (*Source*: TAZ 1997)

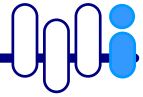
Wenn man die Medizin zur Weltanschauung macht, ja. (*Source*: TAZ 1997)

Collocations



- calculation of significant collocations for ***all entries*** in the corpus
- storage of ***collocation sets*** for simple word forms as well as extracted phrases (like proper names)
- ***sentence-based*** as well as ***immediate-neighbour*** collocations
  - left neighbours
  - right neighbours
- significance measure: G-Test (comparable to log-likelihood measure, see below)

Some Collocation Measures



calculation of the significance of common appearance of two words A, B:

n_A, n_B number of sentences containing A, B

n_{AB} number of sentences containing both, A and B

n_{tot} total number of sentences

Tanimoto-Measure (percentage of double hits with respect to single hits): $sim_T(A,B) = n_{AB} / (n_A+n_B-n_{AB})$

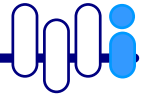
Mutual Information Index (deviance from statistically expected independence) $i(A,B) = \log(n_{AB}\, n_{ges} / (n_A\, n_B))$ $[= \log(p_{AB} / (p_A\, p_B))]$

G-Test (probability of rare events occurring simultaneously)

$x = n_A\, n_B / n_{ges}$

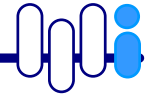
$sig(A,B) = x - n_{AB} \log x + \log n_{AB}!$ (with $2{,}5x < n_{AB}$)

Sentence-Based Collocations for *Retrieval (top 50 only, English Corpus)*



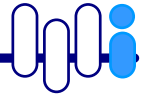
storage (625), text (406), data (390), information (349), search (259), document (211), full-text (204), database (149), Topic (136), indexing (129), software (123), systems (106), documents (103), image (97), CD-ROM (87), optical (82), management (78), Text (75), Verity (71), content-based (71), on-line (62), file (61), capabilities (60), query (60), access (58), processing (58), engine (51), databases (50), electronic (47), Provides (46), archival (46), files (42), hypertext (42), stored (39), archiving (38), users (38), searching (37), Boolean (35), records (35), Gescan (33), applications (33), functions (33), user (33), images (31), queries (30), relational (30), fast (29), searches (29), Information (28), disk (28), Fulcrum (27), ...

Significant *Left* Neighbours of *Retrieval* (English Corpus)



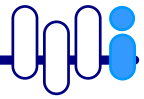
text (401), information (293), data (190), full-text (161), document (96), content-based (77), image (50), Text (44), file (31), Topic (25), concept (20), fast (18), later (18), rapid (18), on-line (16), Concept (15), database (15), quick (15), Full-text (14), easy (13), Information (12), free-text (12), interactive (12), storage (12), message (11), Data (10), computer-assisted (9), subsequent (9), Boolean-based (7), faster (7), news (7), record (7), remote (7), semantic (7), DiscPassage (6), associative (6), index-only (6), allows (5), efficient (5), instant (5), knowledge (5), literary-quote (5), quote (5), sequential (5), Content-based (4), archival (4), concept-based (4), legal-information (4), search (4)

Significant *Right* Neighbours of *Retrieval* (English Corpus)



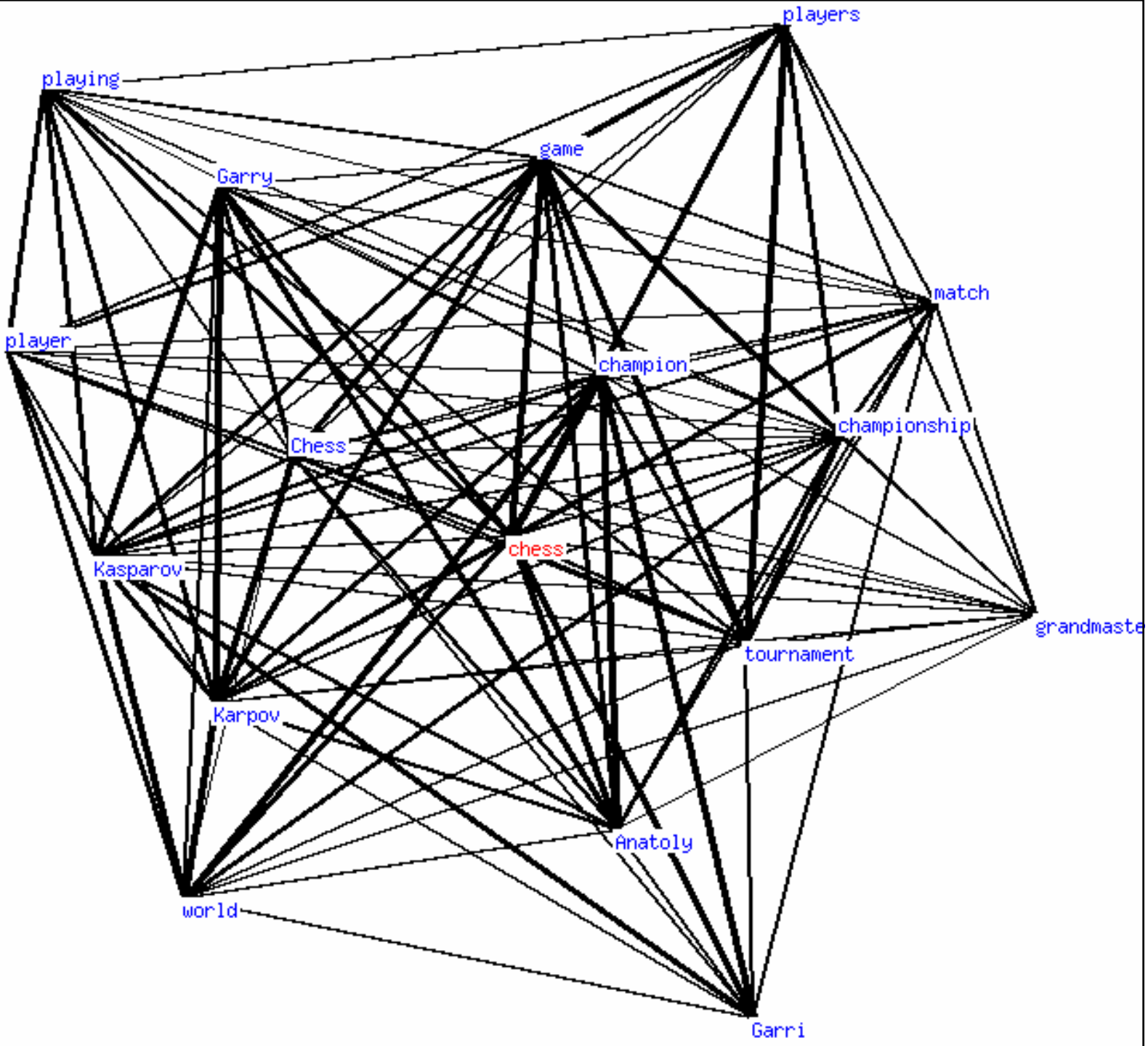
software (175), systems (119), engine (83), capabilities (58), program (31), functions (25), package (23), service (19), services (17), capability (15), tool (13), methods (10), packages (10), process (10), engines (9), programs (9), technology (9), utility (9), stations (8), times (8), clients (7), mechanism (7), operations (7), facilities (6), mechanisms (6), product (6), subsystems (6), techniques (6), time (6), method (5), performance (5), products (5), speed (5), strategies (5), tools (5), client (4), purposes (4), speeds (4)

Visualisation of Collocation Sets



- graph-based visualisation of collocation sets
- graph drawing based on *simulated annealing*
- real time drawing, i. e. for every word with a minimum number of significant collocations a graph may be drawn
- interactive graphs, i. e. for each node in the graph, a new graph may be selected interactively
- www access

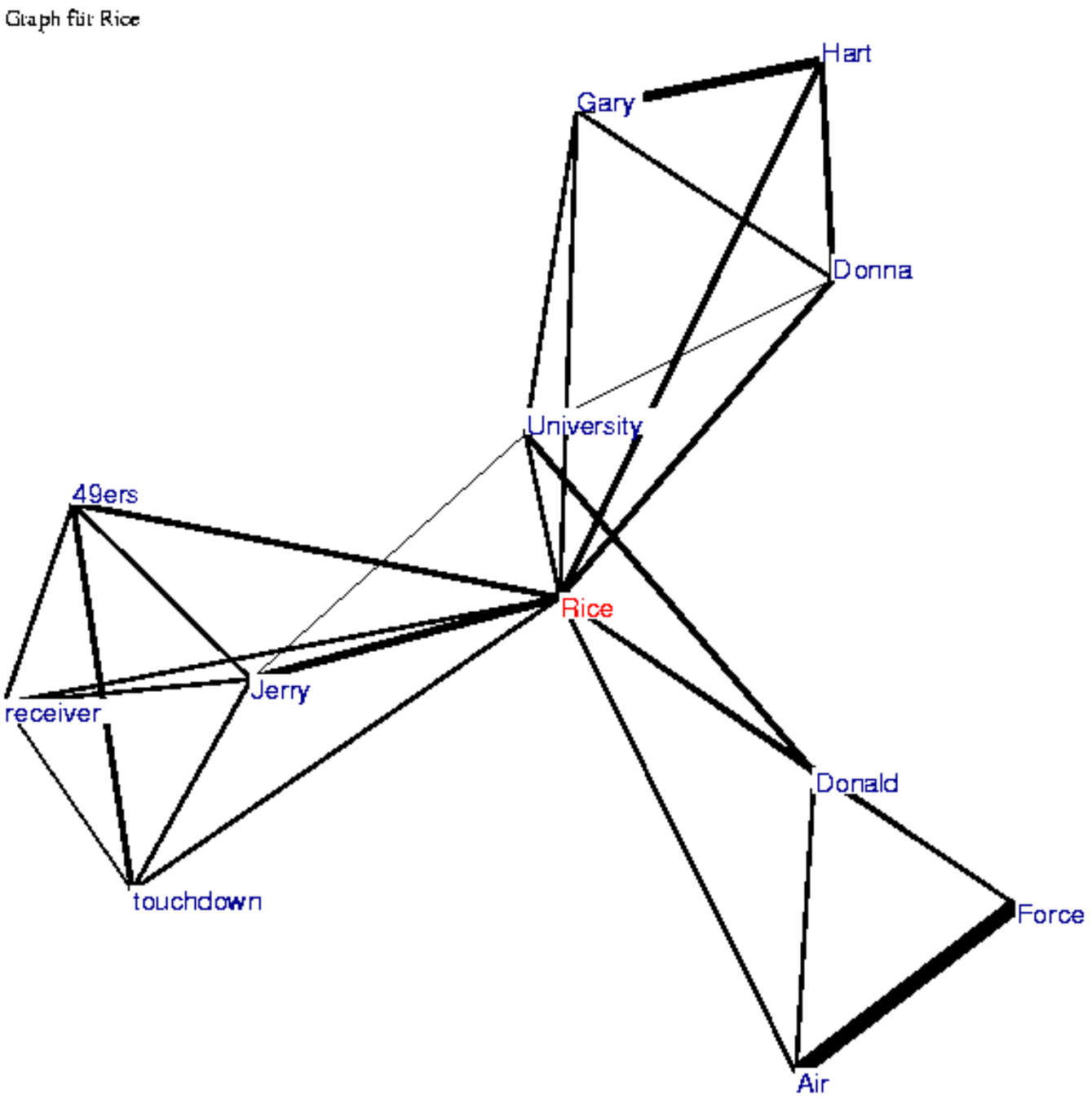
Chess


players
playing
game
Garry
match
player
champion
championship
Chess
chess
Kasparov
grandmaste
tournament
Karpov
Anatoly
world
Garri

Separation of Meanings

- different meanings can be visualised as subsets of a larger set of collocation

- Example: **Rice**
  - Rice University
  - Donna Rice
  - Jerry Rice
  - Rice Air Force Base

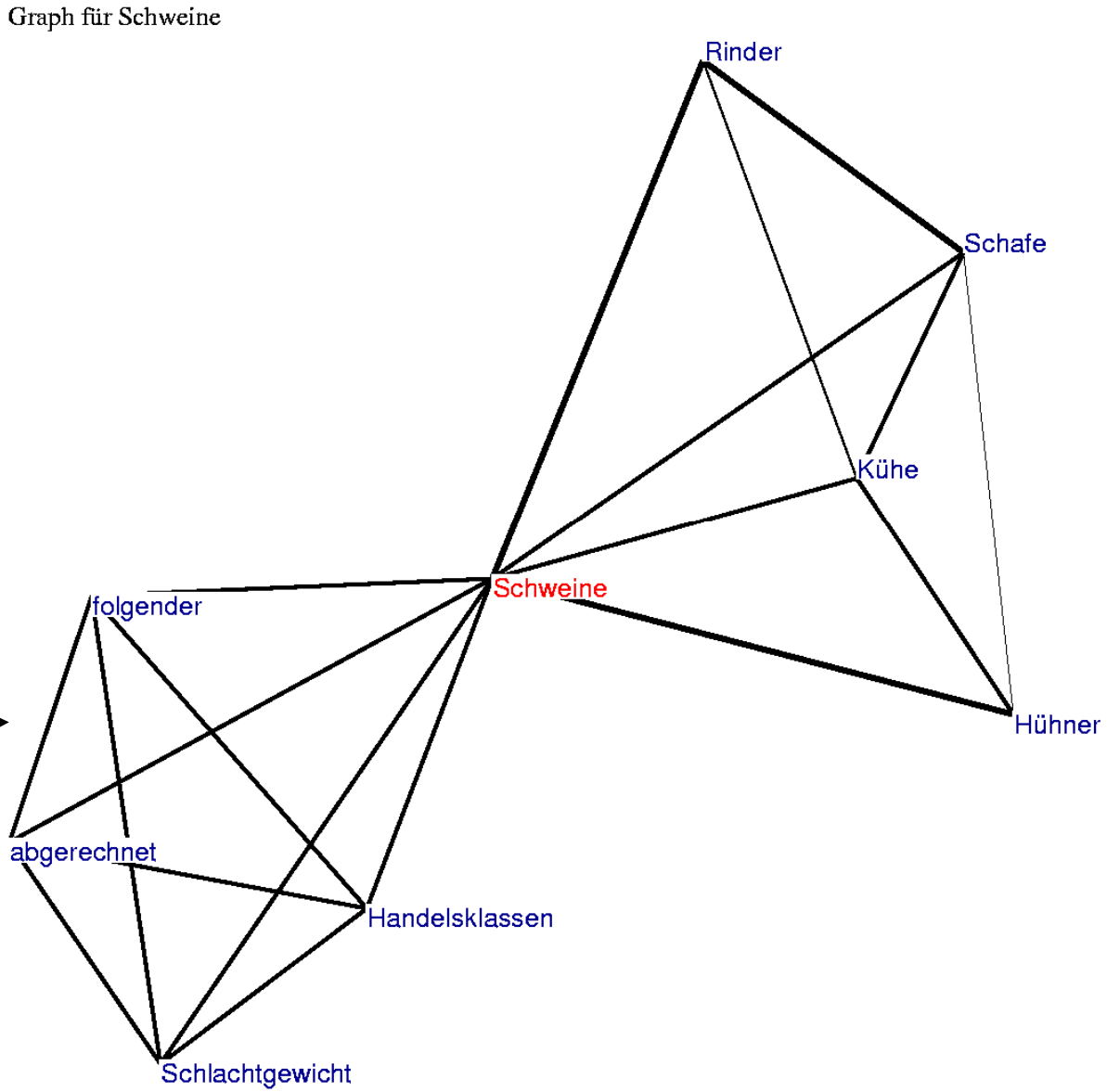


„Schweine“ (świni)

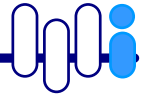
cohyponyms

- two meanings:
  - pigs as a kind of animal like cows, sheep etc.
  - pigs as merchandise which are slaughtered, sold etc.

technical Concepts

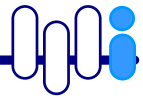


Time-sliced corpora



- Collection of corpora of **equal size** (and from equal sources) for
  - **different years** since 1995
  - **every day** of the year (since 2002)
- Sources
  - online newspapers and journals
  - collection during the night, analysis in the morning
- Selection criteria
  - daily availability
  - broad coverage of topics
  - automated collection of texts

Processing and Analysis of Day-based Corpora

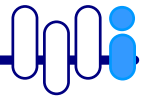


1. Availability of large reference corpus (size factor ~1000)
2. segmentation (sentences, words)
3. indexing and counting
4. calculation of sentence and immediate neighbohood collocations
5. storage in relational database (MySql)

Typical Results for a Day-based Corpus

Sentences	17645
Number of tokens	255680
Number of types (inflected word forms)	41031
Number of sentence collocations (word pairs)	90528
Number of immediate neighborhood collocations (word pairs)	6581
Neighborhood collocations of capitalized words	546
Relative corpus size in comparison with reference corpus	1 : 936,16

Selection of “Word of the Day” candidates



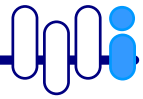
- definition of threshold values (relative frequency etc., see below)
- combination of parameters
- error correction (e.g. typos)

- Result: List of candidates without categorization

Candidate Selection Parameters

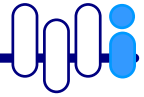
Parameter	**Motivation**	Heuristics
n_1 Frequency in day-based corpus	Relatively frequent words may indicate important topic	$n_1 > 8$
n_{ges} Absolute frequency in reference corpus	A baseline indicates that the concept is known	$n_{ges} > 20$
1000 n_1 / n_{ges} relative weighting of term	Indicator for a frequency that is higher than expected	1000 $n_1 / n_{ges} > 16$ (“1000”: normalization of corpus size)

Sample candidate selection (June 2002)



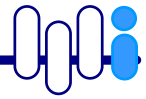
Abschlussbericht, Abu, Annecy, Babcock, Barrichello, Bild am Sonntag, Bildungspolitik, Bin Laden, Bondy, Bulmahn, Bundespolitik, Butler, Coast, Coulthard, Daglfing, Djerba, Elfmeterschießen, Fifa, Foul, France Télécom, Fritz Walter, Fußball-WM, Gläubigerbanken, Golden Goal, Großaktionäre, Guus Hiddink, Gymnasien, Hartz, Hewitt, Hiddink, Ilhan, Insolvenz, Kahn, Koreaner, Kroetz, Lifestyle, Lukaschenko, Manfred Stolpe, Matthias Platzeck, Medienberichten, Mobilcom, MobilCom, Montoya, Müntefering, Naturwissenschaften, Nürburgring, NZZ, Oliver Kahn, Pisa, Plank, Platzeck, Platzecks, Ralf Schumacher, Rößler, Rubens, Rubens Barrichello, Rüge, Schill, Schmid, Schulsystem, Schuss, Senegal, Senegals, Sevilla, Skibbe, Stage, Stol-pe, Stolpes, Strüver, Südkorea, Südkoreaner, Südkoreas, Talkline, Tokyo, Trienekens, Völler, Walküre, Wienand, Wittenberge, WM-Halbfinale, Zeppelin, Zuwanderungsgesetz

Categorization



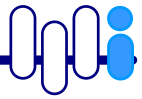
- simple set of category buckets (like in a newspaper)
- automated extraction of candidate lists
- manual post-processing of new terms (web-based interface)
- growing database of relevant and categorized terms

Result Visualization



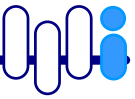
- Lists of categorized “words of the day”
- hyperlinks to the original sources
- collocation graphs for
  - day-by-day contexts
  - comparison with reference corpus
- line-chart visualization of trends (frequency of terms / days for which a term was selected as “word of the day”)

Words of the Day (22. 6. 2003)



	Wortschatz : Wörter des Tages : 22.06.2003	
sportowiec, trener, działać sport	Sportler, Trainer, Funktionäre	Lennox Lewis · Vitali Klitschko
	Sport	Etappe · Klitschko · NBA
polityczny	Politiker	Bundeskanzler Gerhard Schröder · Eichel · Friedman · Merkel · Michel Friedman · Saddam · Seehofer · Wowereit
organizacja	Organisation	BMW · Bundesrat · IG Metall · Konvent · Union · Volvo · WestLB
zdarzenie	Ereignis	EU-Gipfel · Regensburg · Streik
hasło	Schlagwort	Arbeitskampf · Gesundheitspolitik · Kokain · Massenvernichtungswaffen · Porto · Videoüberwachung · Wehrpflicht · Zahnersatz · Zuwanderung
miejsce	Ort	Afghanistan · Irak · Iran · Kabul · Russland · Teheran · Thessaloniki
osoba (sztuka, kultura, nauka)	Personen aus Kunst, Kultur und Wissenschaft	
osoba (rozmaity)	sonstige Personen	Ermittler · Hacker · Harry Potter · Lewis
	««21.06.2003«« Wörter des Tages	

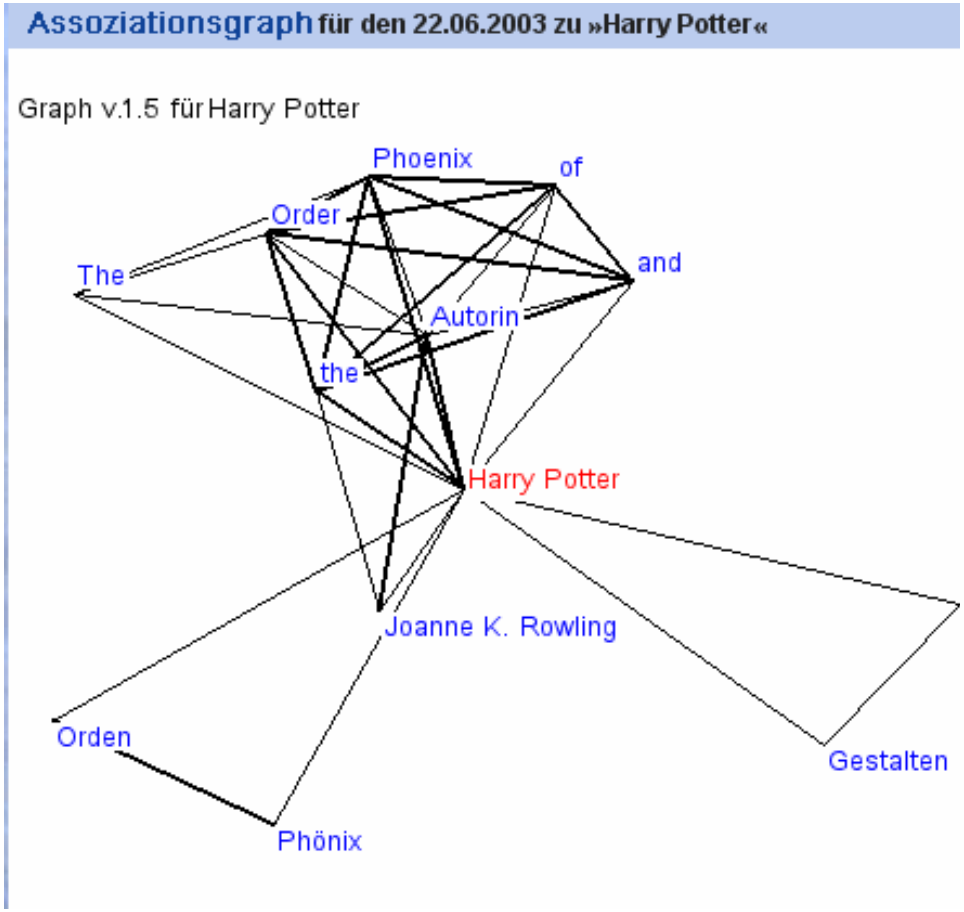
Words of the Day (19. 4. 2006)



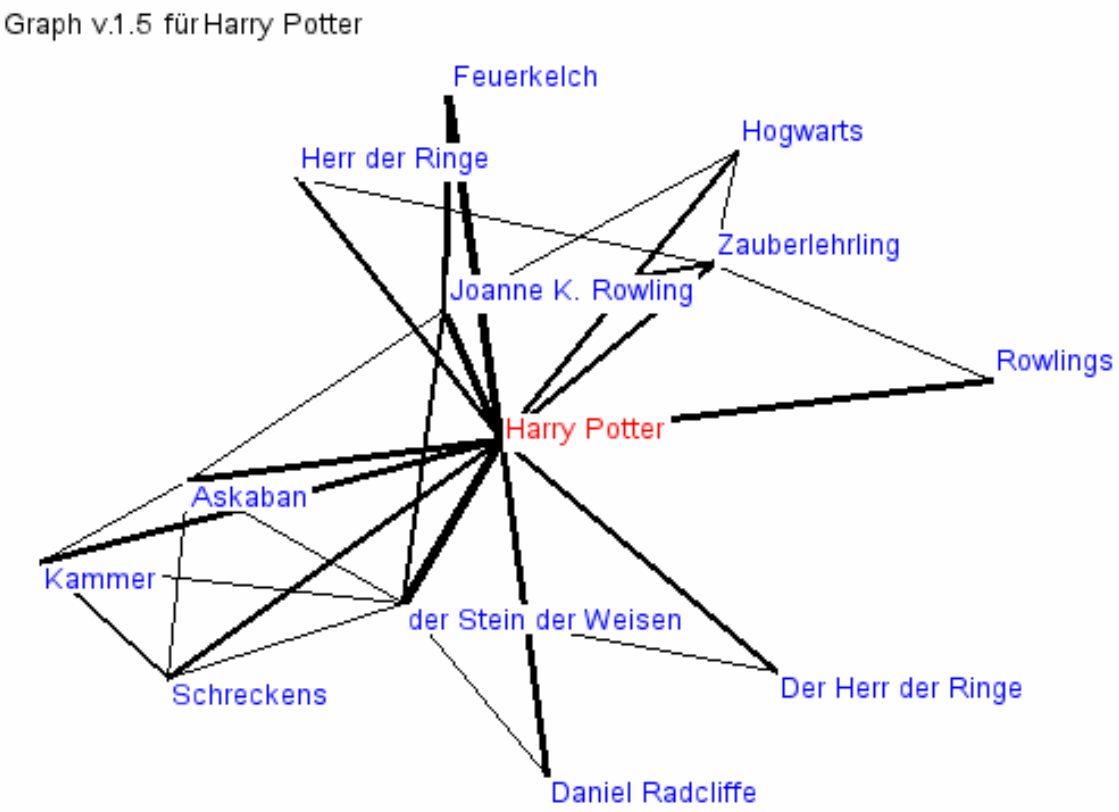
sportowiec, trener, działać sport	Armin Veh · Drogba · Inzaghi · Ismael · Jens Lehmann · Klitschko · Landgraf · McCormack · Meijer · Montoya · Rijkaard · Ronaldinho · Veh · Wladimir Klitschko · Wosz
	AC Mailand · Alemannia Aachen · Eisbären · FC Barcelona · Fußball-Regionalliga · Fußball-Weltmeisterschaft · Marathon · Milan · NBA · NHL · Sportdirektor · VfL Bochum · Werfer
polityczny	Bosbach · Brandt · Burns · Ciampi · Dieter Wiefelspütz · Hartz · Hu Jintao · Innenminister Jörg Schönbohm · Jens Bullerjahn · Mahmud Abbas · Ministerpräsident Ehud Olmert · Ministerpräsident Wolfgan Böhmer · Nehm · Präsident Mahmud Ahmadinedschad · Romano Prodi · Rumsfeld · US-Präsident George Bush
organizacja	AMG · Air Berlin · Al-Aksa-Brigaden · Autonomiebehörde · BKK · BaFin · DFL · DHL · GE · Gasprom · Haaretz · Hamas · Hamas-Bewegung · IAEO · IG Metall · Islamischer Dschihad · KaDeWe · Kölnische Rundschau · Lenovo · Linkspartei · Opec · Oracle · Philips · TNK-BP · Tchibo · Times · UN-Sicherheitsrat · WASG · ZEW
zdarzenie	Anschlag · Attentate · Beben · Hochwasser · Hurrikan · Ostermontag · Ostern · Ostersonntag · Osterwochenende · Rast · Selbstmordanschlag · Tarifabschluss · Tarifverhandlungen · Terroranschlag · Verhandlungsrunde
hasło	Akupunktur · Anreicherung · Arbeitsbedingungen · Arbeitskampf · Arbeitslosengeld II · Atomanlagen · Atomenergie · Atomkonflikt · Atomprogramm · Atomstreit · Benzinpreise · Bestechlichkeit · Deich · Dschihad · Ehrenmord · Einmalzahlungen · Erdöl · Hamas-Regierung · Katie · Katrina · Kernenergie · Koalitionsvertrag · Krankenstand · Königsklasse · Linux · Pulitzerpreis · Rohöl · Sonntagmorgen · Sorgerecht · Steuererhöhungen · Stimmzettel · Sudoku · Tarifkonflikt · Tarifrunde · Tarifstreit · US-Ölpreis · Urabstimmung · Uran · Vogelgrippe · Vätermonate · Warnstreik · Wasserstand
miejsce	Aachen · Alexandria · Bamberg · Boston · Chelsea · Dallas · Deggendorf · Donau · Elfenbeinküste · Eschborn · Flensburg · Gibraltar · Gummersbach · Imola · Iran · Katar · New Orleans · Petersdom · Reutlingen · Romano · San Antonio · San Francisco · Teheran · Tel Aviv · Tschernobyl · Weißrussland
osoba (sztuka, kultura, nauka)	Björk · Ellison · Michael Jäger · Tom Cruise
osoba (rozmaity)	Bauermann · Chris · Französin · Hatun Sürücü · Hooligans · Hunold · Kopten · Lynch · McCarthy · Pischetsrieder · Popstar · Selbstmordattentäter · Söder · UN-Botschafter · Wetzel · Zacarias Moussaoui

Compariosn of Collocation Graphs („Harry Potter “, 22. 6. 2003 / Reference Corpus)

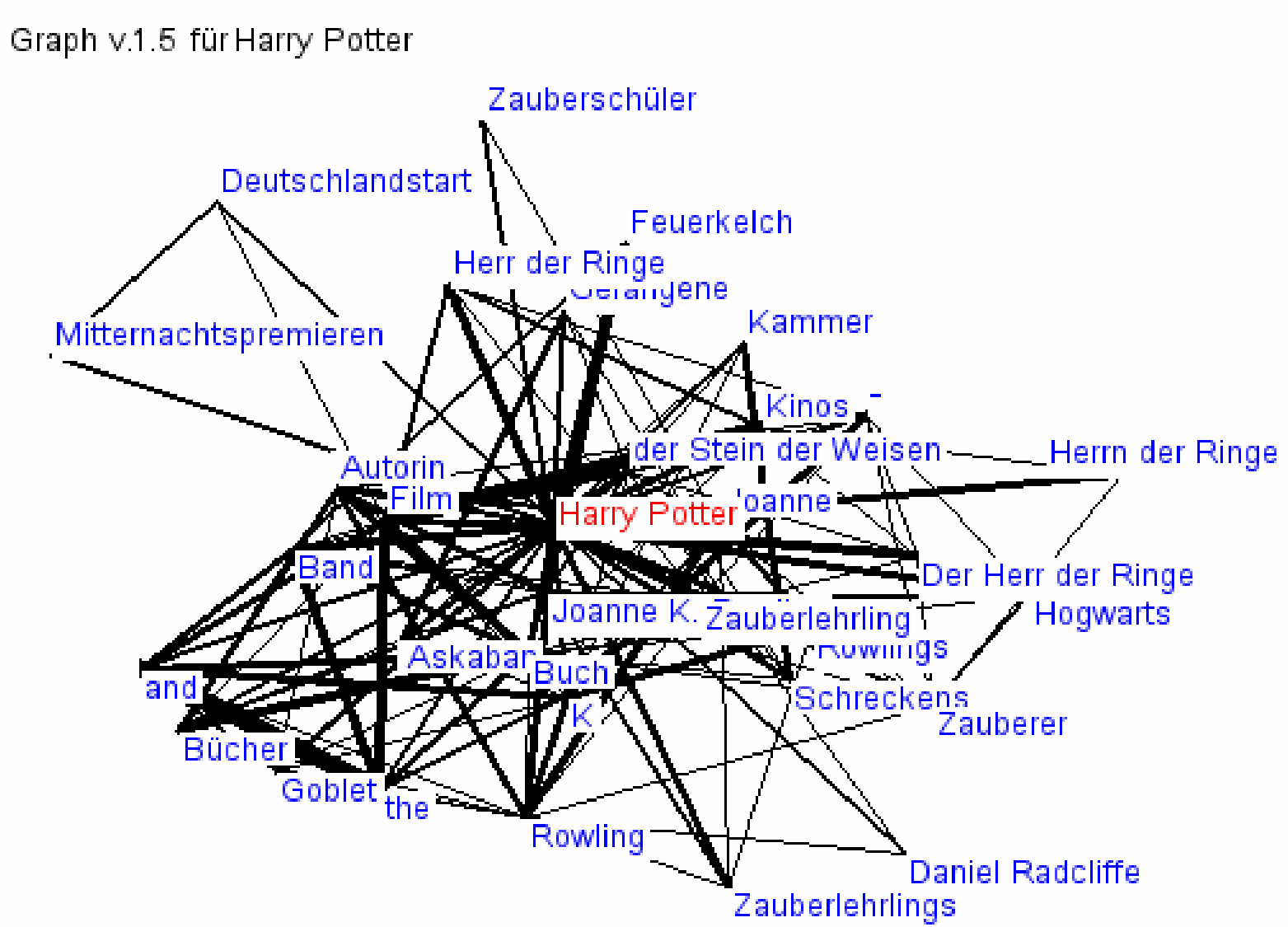
Collocation Graph 22.6.2003



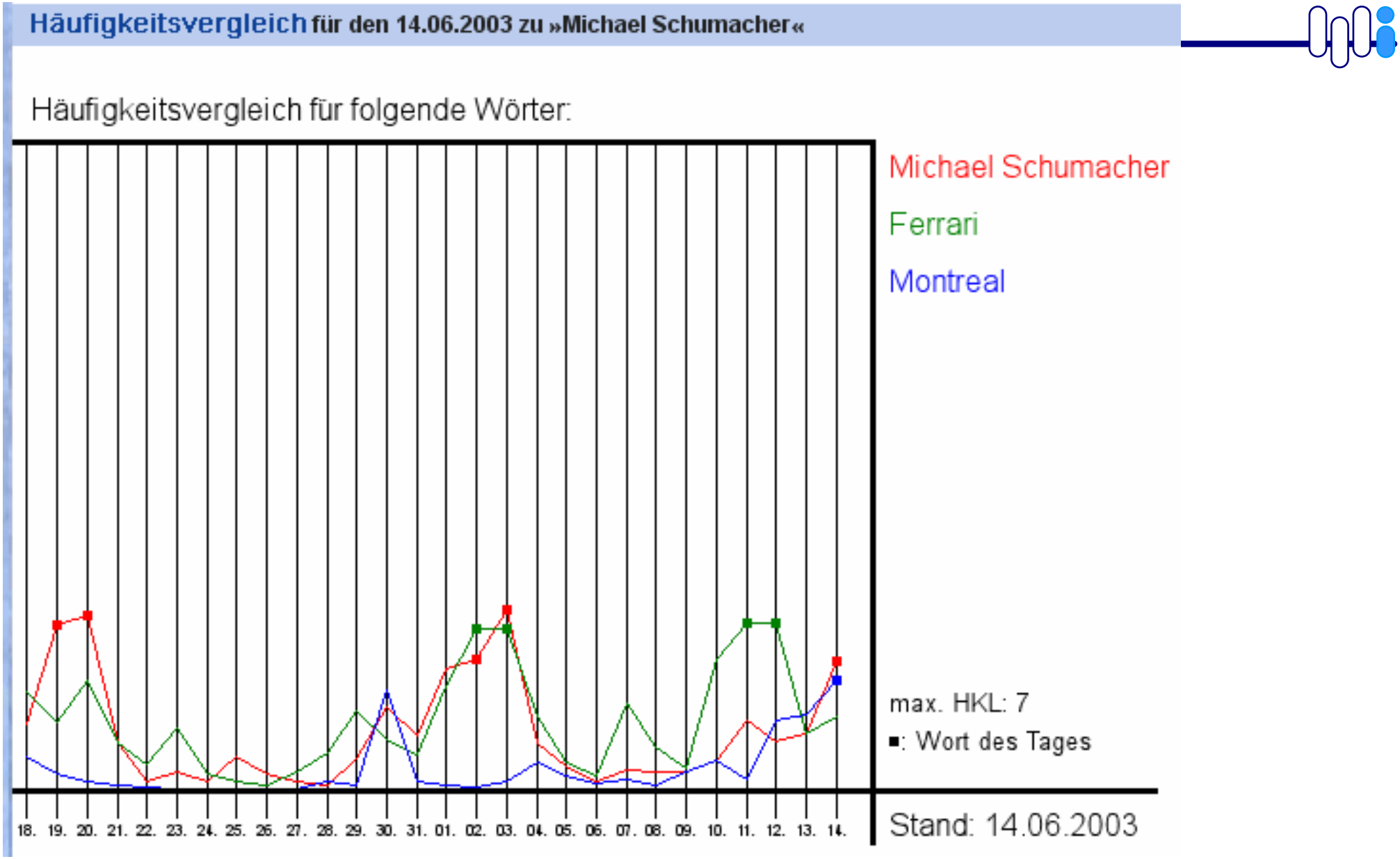
Collocation Graph in Reference Corpus



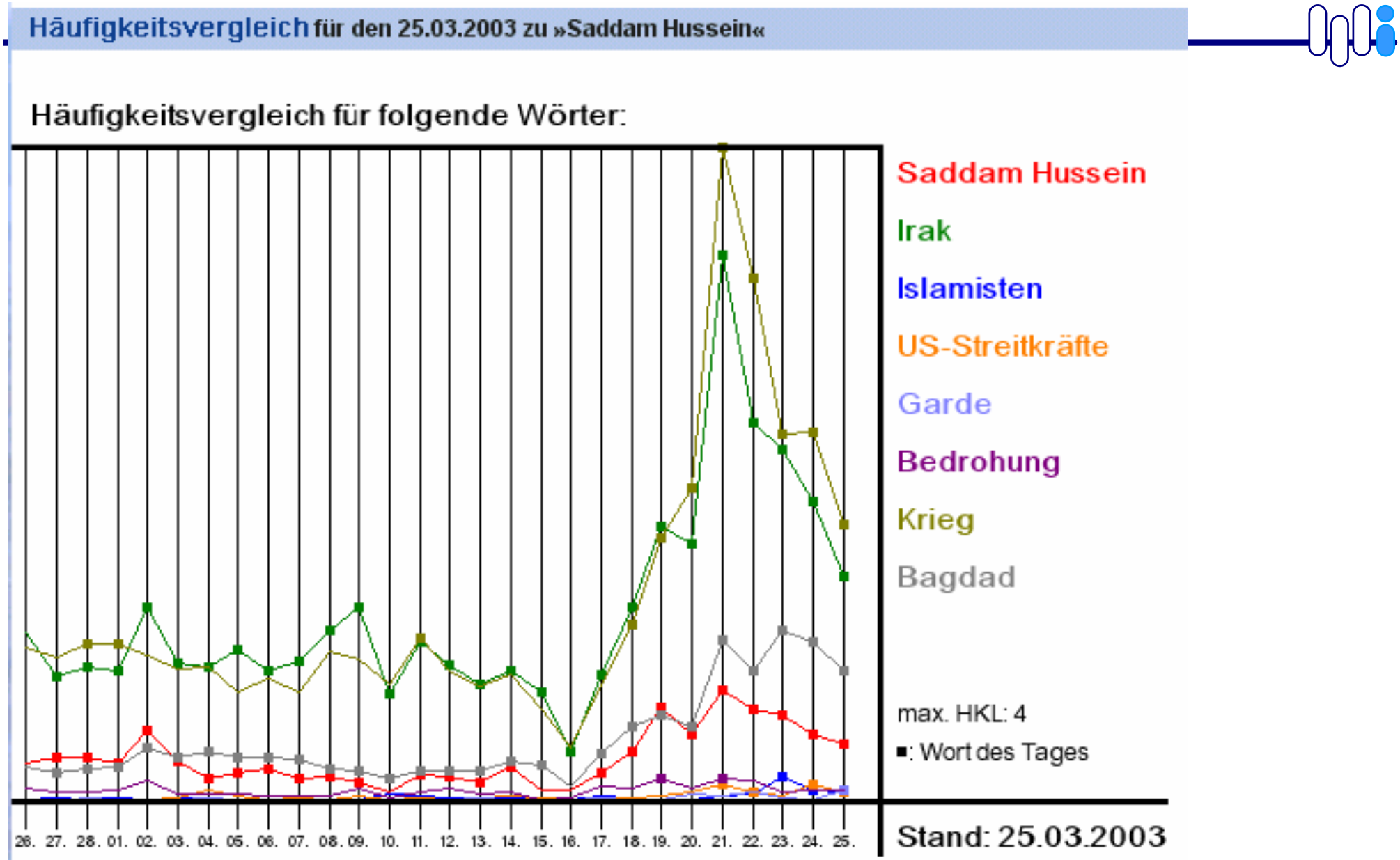
Collocation Graph in Reference Corpus (expanded, with more concepts shown)



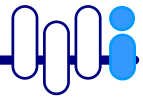
Visualization of Activity Trends: Michael Schumacher (14. June 2003)



Visualization of Activity Trends: Saddam Hussein (25. 3. 2003)

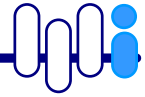


Open Problems



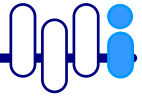
- source collection must be adapted frequently due to changes in websites
- inflected forms might be aggregated (stemming)
- duplications of terms of multiword terms (e.g. for proper nouns: Rubens, Rubens Barrichello)
- source selection might be standardized by using media usage data (by how many people has a certain text been read – socio-linguistic aspects)
- development of adequate media usage indices
- cross-media collection of data, e.g. by speech recognition of spoken radio and TV data

Possible Applications



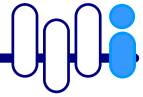
- media analysis: media usage and impact
- trend research
- lexicography (collocations, neologisms, relevant associations)
- taking into account knowledge resources like WordNet

Time-based comparison of collocation sets



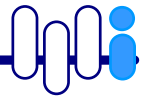
- Idea: comparing collocation sets in corpora taken from different time slices
- Media analysis or language change?

Problems



- comparison metrics
  - set-based (e.g. intersection or Hausdorff metrics for set similarity)
  - rank-based (rank correlation coefficients)
  - function analysis for time series (?)
- adequacy of time slices (day, month, year, decade, ...)
- availability of
  - corpus data (e. g. non-textual data)
  - linguistic annotation
- selection of interesting concepts
  - classifying types of change
- result interpretation
- going from pragmatics to semantics to syntax???

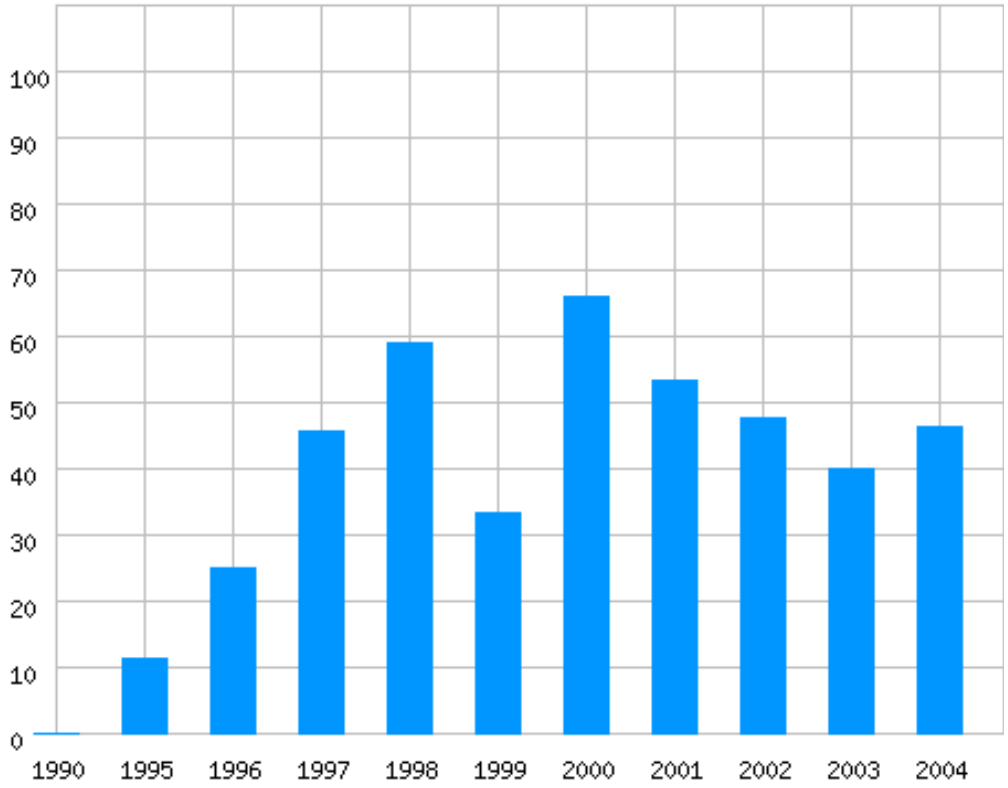
Preliminary results of analysis



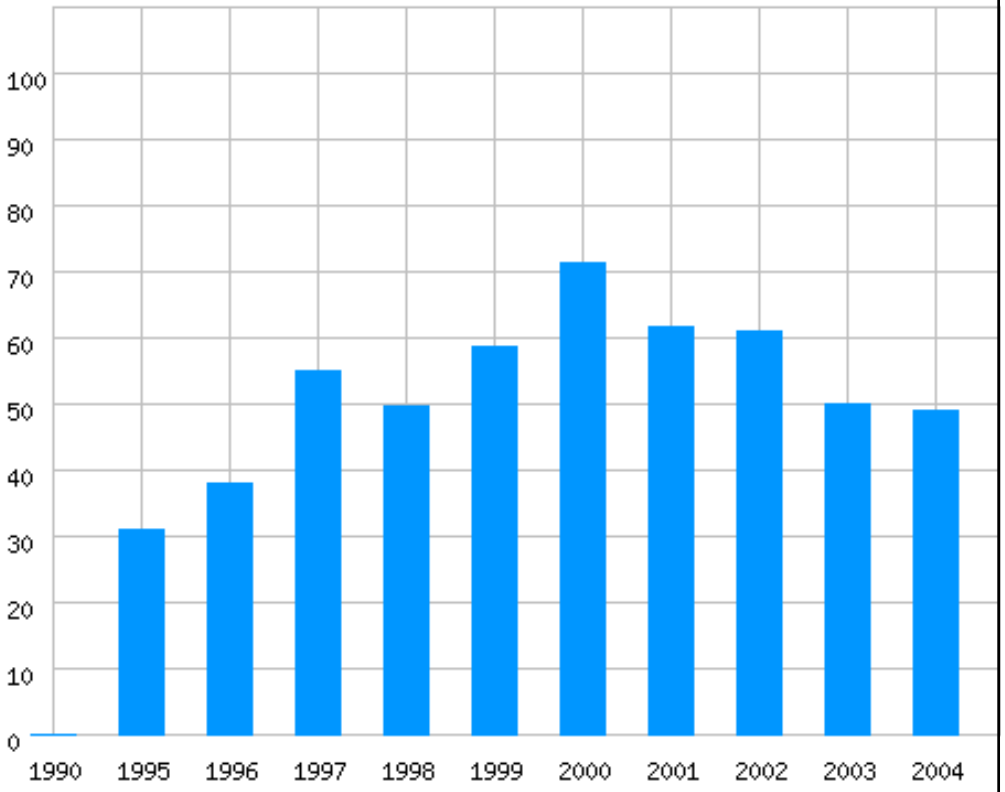
- comparison of collocation sets for selected words (nouns, verbs, adjectives, ...)
- criteria
  - immediate neighborhood vs. sentence
  - monthly vs. yearly data
  - rate of new collocates compared with accumulated predecessing periods versus absolute change in comparison to preceding period

Bush

percentage of new immediate neighborhood collocations compared with accumulated collocations from preceding years

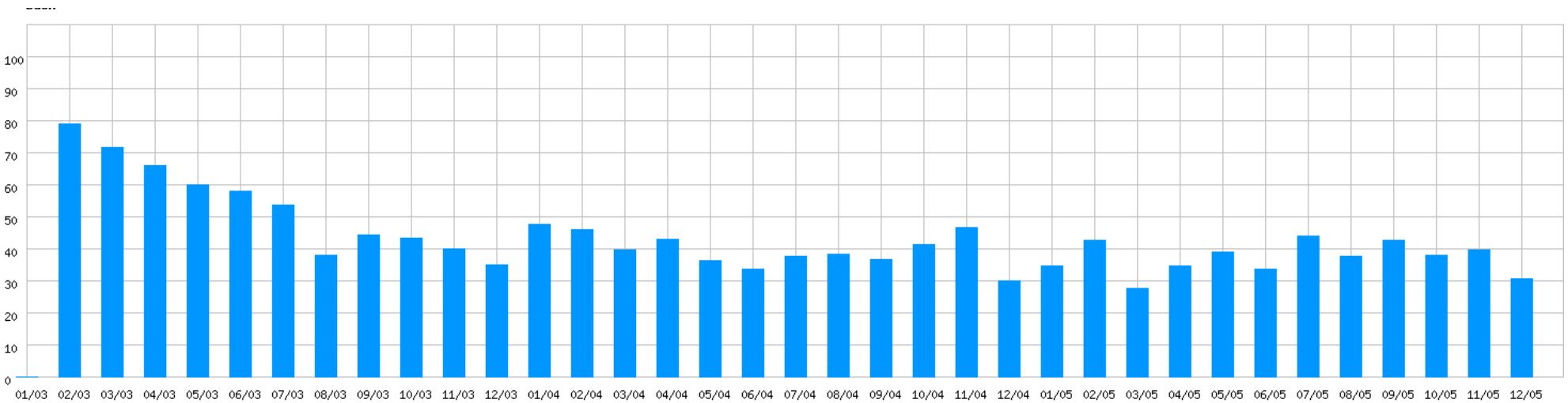


percentage of new sentence collocations compared with accumulated collocations from preceding years



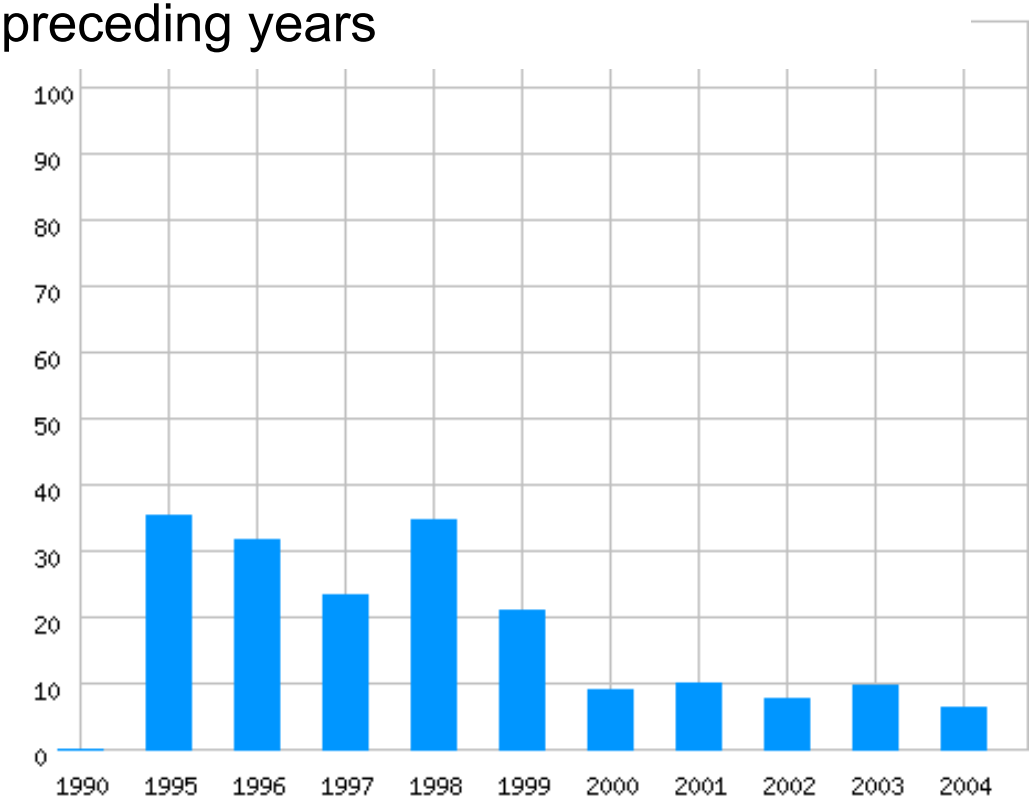
Bush – monthly data

percentage of new ***sentence collocations*** compared with accumulated collocations from preceding **months**

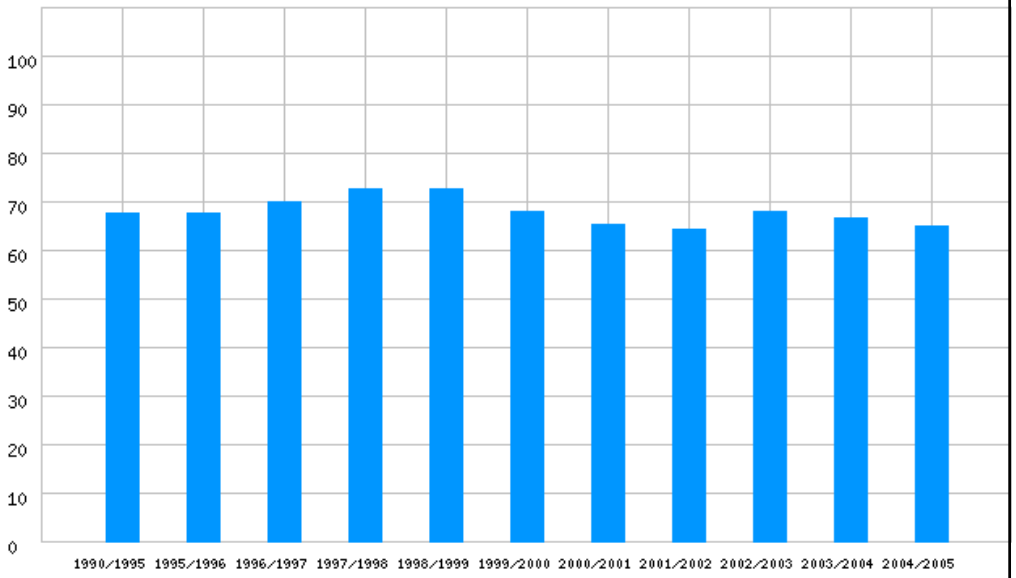


Chance (chance – sposobność)

percentage of new ***immediate neighborhood*** collocations compared with accumulated collocations from preceding years

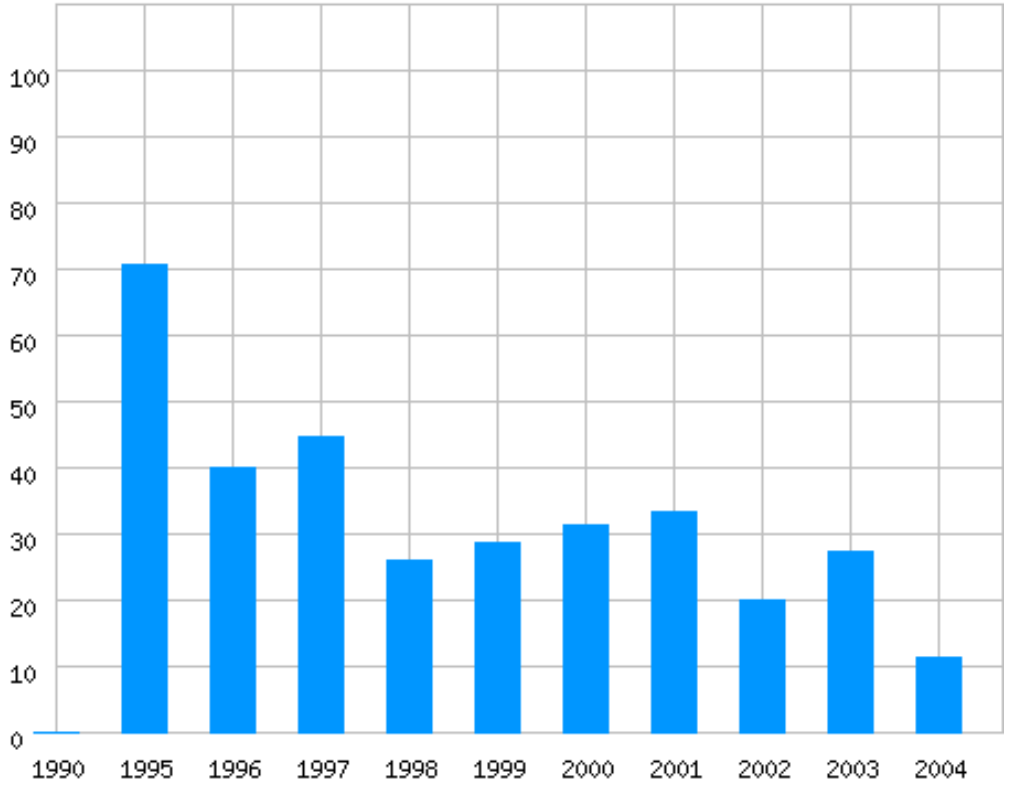


percentage of different ***immediate*** neighborhood collocations in biannual comparison (union –intersection)

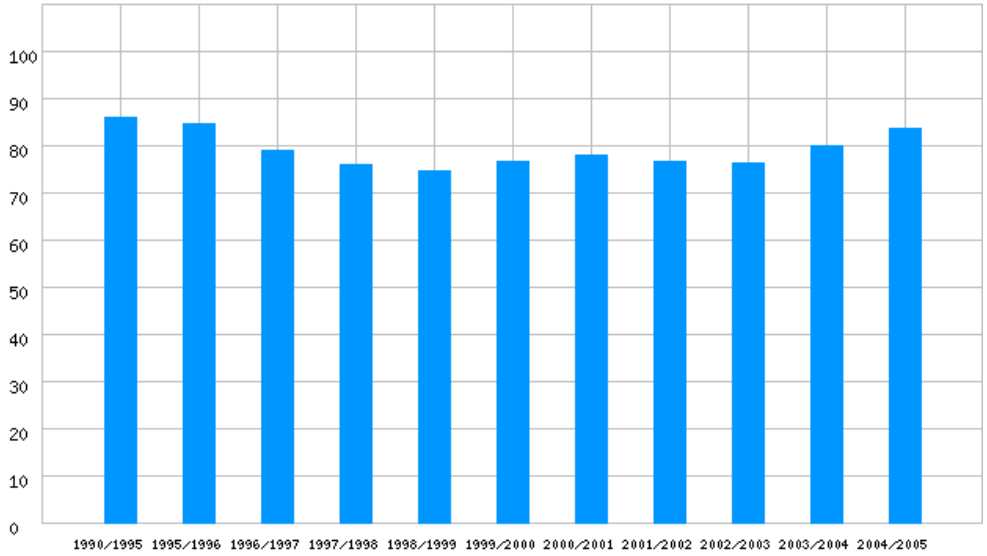


Erinnerung (wspomnienie)

percentage of new ***immediate neighborhood*** collocations compared with accumulated collocations from preceding years

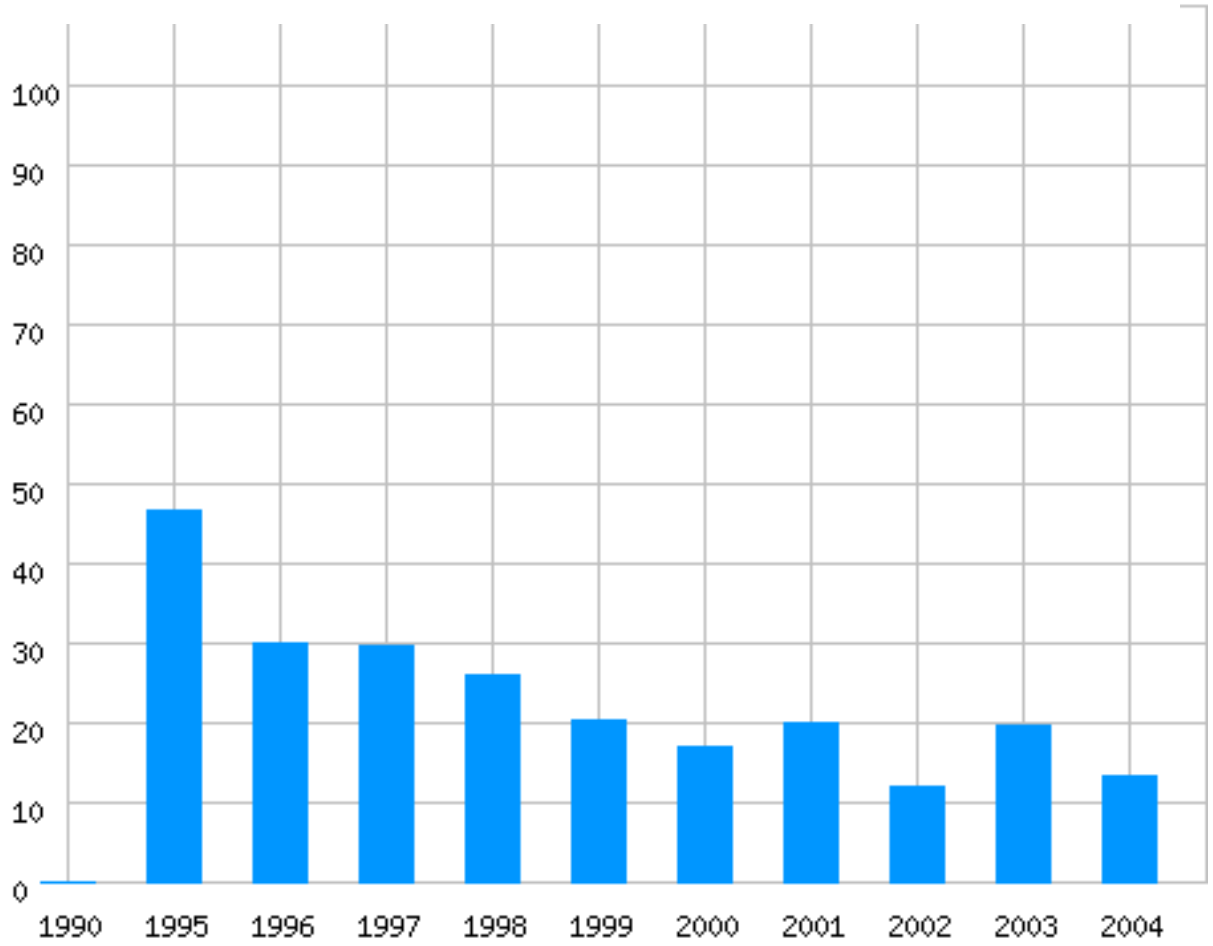


percentage of different ***immediate*** neighborhood collocations in biannual comparison (union –intersection)



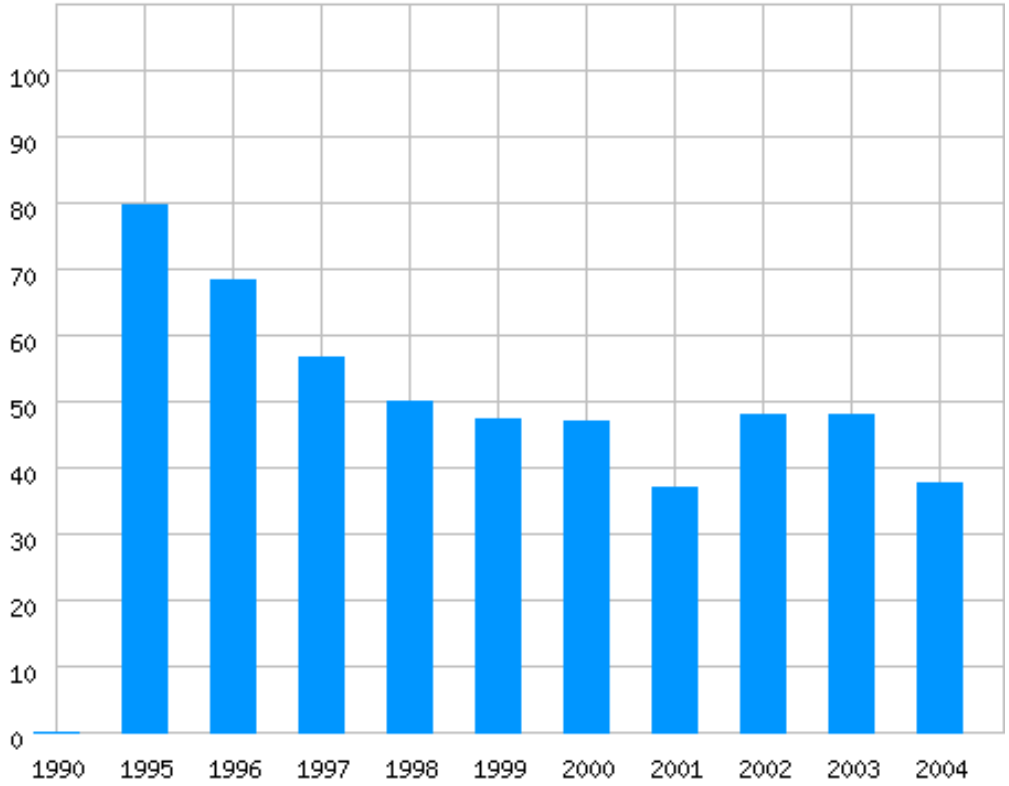
Leben (life – życie)

percentage of new ***immediate neighborhood*** collocations compared with accumulated collocations from preceding years

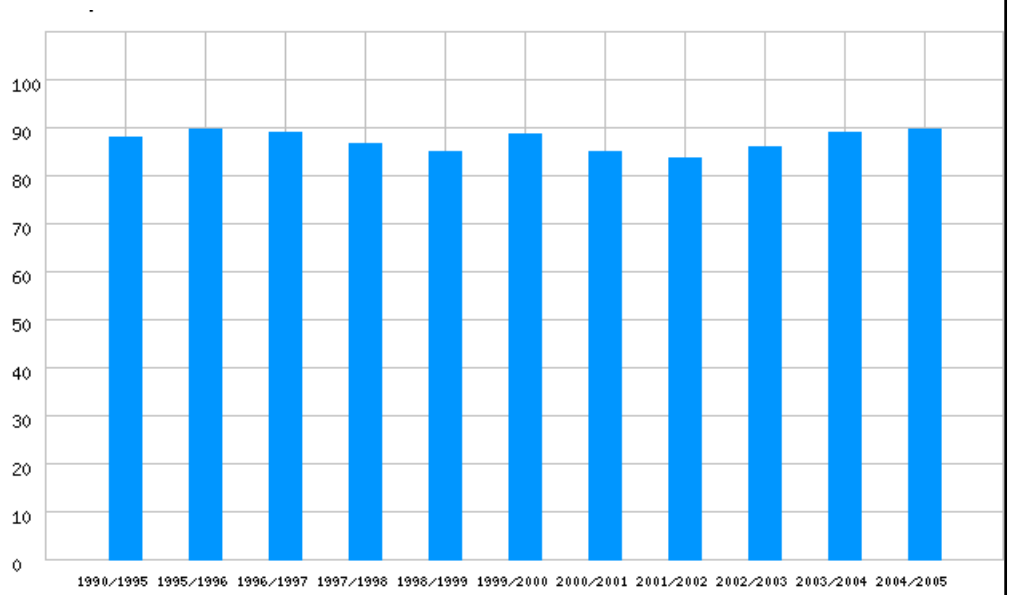


kämpfen (to fight - walczyć)

percentage of new ***immediate neighborhood*** collocations compared with accumulated collocations from preceding years

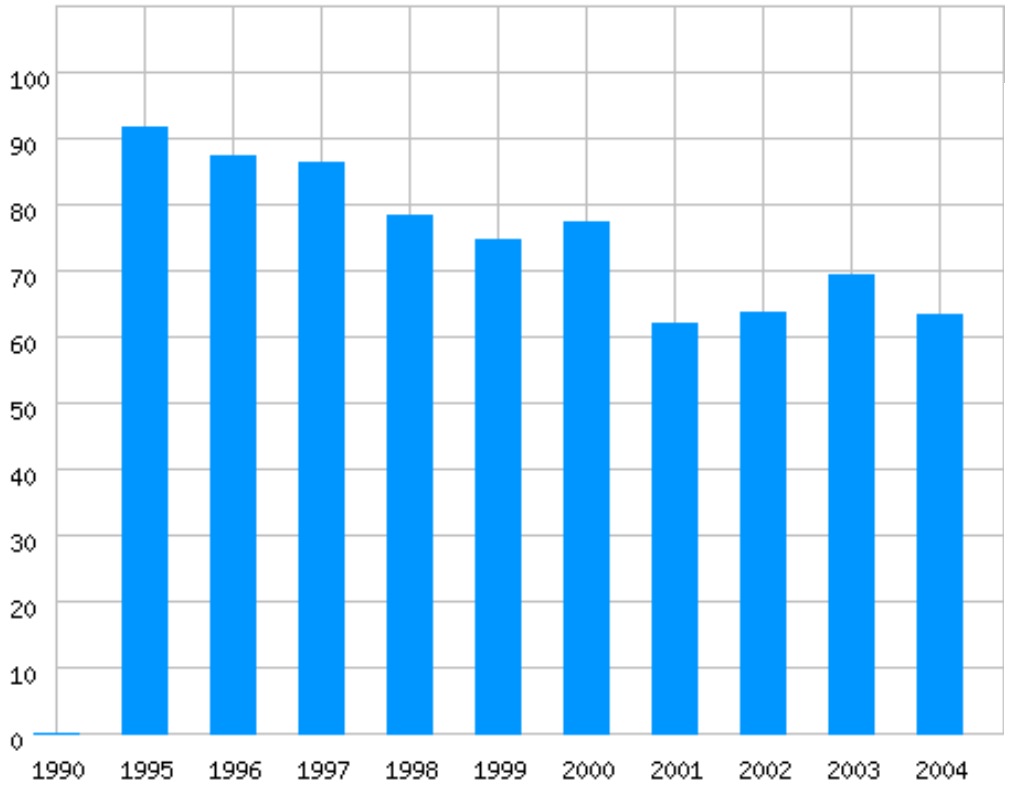


percentage of different ***immediate*** neighborhood collocations in biannual comparison (union –intersection)

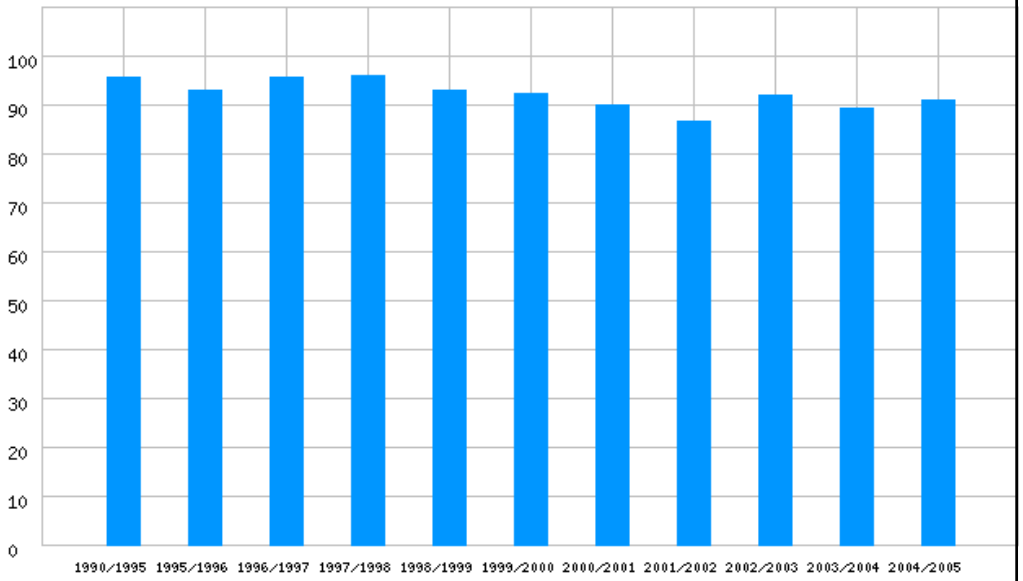


traditionell (traditional – tradycyjny)

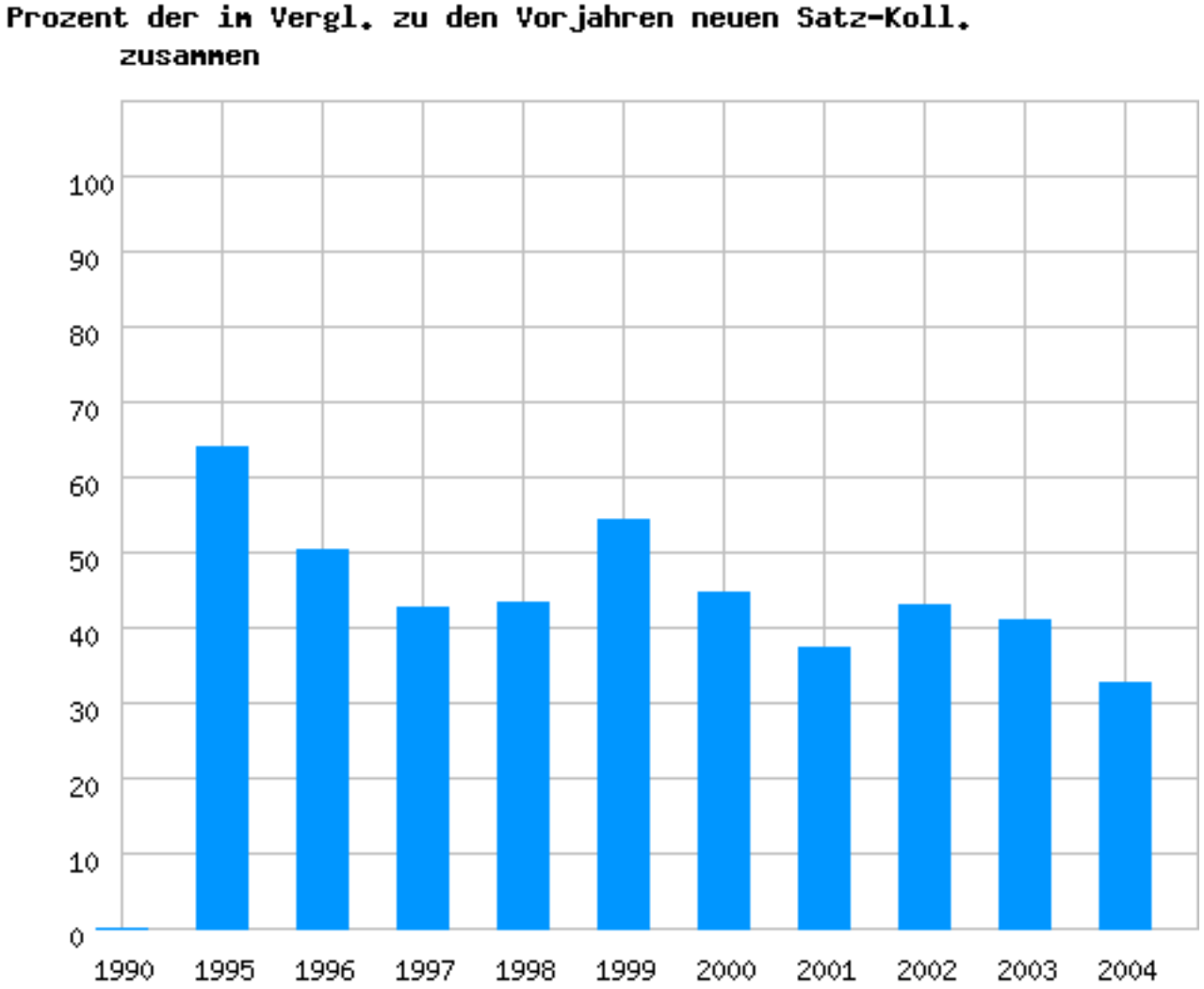
percentage of new ***immediate neighborhood*** collocations compared with accumulated collocations from preceding years



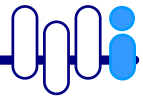
percentage of different ***immediate*** neighborhood collocations in biannual comparison (union –intersection)



zusammen (together – razem)



Outlook – next steps



- not enough monthly data for most words
- analysis for larger time slices (e.g. decades taken from DWDS)
- cleansing / error correction
- calculation of curves for all words / concepts in the database
- finding changing types for different categories (proper nouns / common nouns, verbs, adjectives)
- finding additional change indicators like changes in relative frequency
- in-depth analysis of collocation sets