DeepDict

A Graphical Corpus-based Dictionary of Word Relations

Eckhard Bick
University of Southern Denmark & GrammarSoft ApS

Motivation – the lexicographer's view

- a) lexicography: better data
  - corpus data have better coverage and better legitimacy than introspection or chance quotes from literature
  - Quantity: the more the better
  - Authenticity: real running data from mixed sources (Internet), or at least a source with mixed topics (Wikepedia, news text)
  - given a corpus, a grammatically annotated one is best for the extraction of lexical patterns and statistics
  - lexical patterns are best based on linguistic relations (subject, object etc.) rather than mere adjacency in text
  - even given a corpus that satisfies all of the above, it is cumbersome to search for and quantify lexical patterns
  - even a statistics-integrating interface like our CorpusEye will only provide data for one pattern at a time

Motivation – the user's view

- b) lexical information: better accessibility
  - electronic vs. paper: no size limitations, easy searching, “depth-on-demand” (QuickDict vs. DeepDict)
  - passive (“definitional”) vs. active (“productive-contextual”)
  - Advanced Learner's Dictionary: information on how to use a word in context – syntactic and semantic restrictions and combinatorial aspects, e.g. A gives x to B (A,B = +HUM, x,y = -HUM)
  - “live” examples
  - but how to show, for a given entry word, all constructions and examples, and how not to forget any?

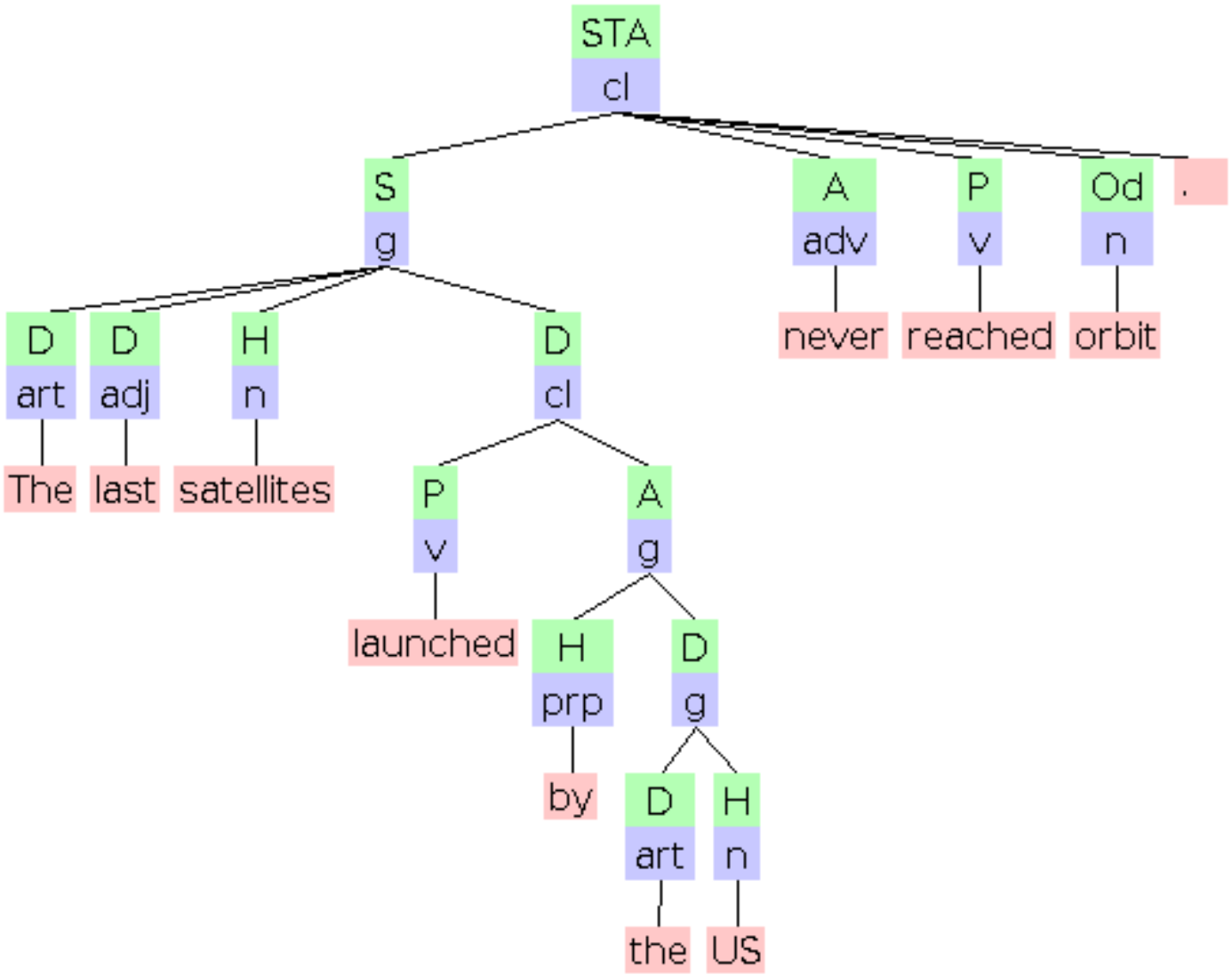
Idea: graphical presentation of **lexical complements** based on **dependency statistics**

- annotate a corpus
  - Peter [Peter] @SUBJ ate a couple of apples [“apple”] @ACC
  - Cats [“cat”] @SUBJ eat mice [“mouse”] @ACC
- generalize by collecting and counting dependency pairs of lemmas (simplified):
  - PROP_SUBJ -> eat, cat_SUBJ -> eat
  - apple_ACC -> eat, mouse_ACC -> eat
- present the result in list form
  - {PROP,cat} SUBJ -> eat <- {apple,mouse} ACC

Dependency tree annotation

The <def>	ART	@>N	#1->3
last <num-ord>	ADJ	@>N	#2->3
satellites <Vair>	N P NOM	@SUBJ>	#3->9
launched	V PCP2 PAS	@ICL-N<	#4->3
by	PRP	@<PASS	#5->4
the <def>	ART	@>N	#6->7
US <civ>	PROP F S	@P<	#7->5
never <atemp>	ADV	@ADVL>	#8->9
reached	V PAST	@FMV	#9->0
orbit <L>	N S NOM	@<ACC	#10->9
$.			#11->0

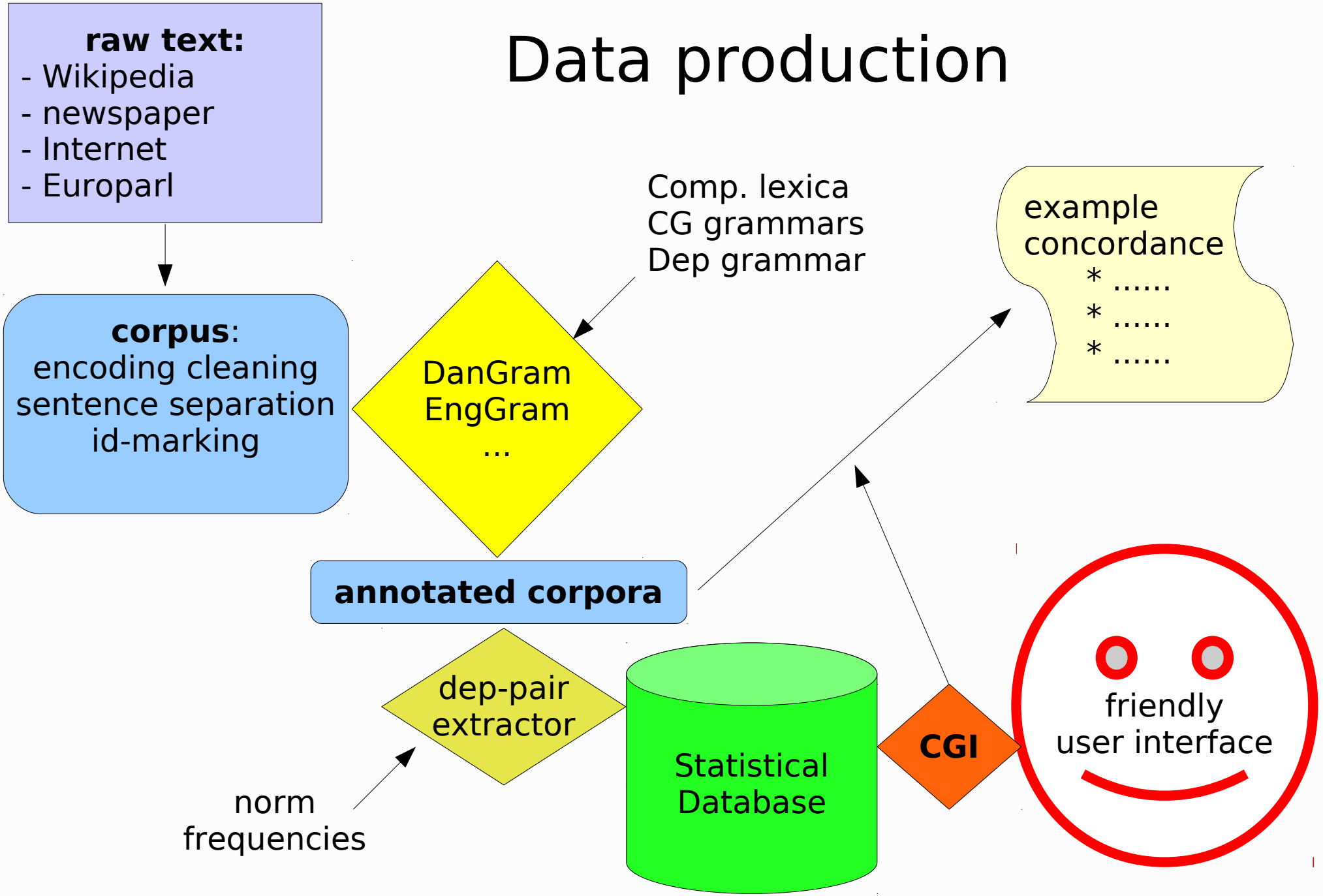
equivalent constituent tree



how to distinguish between typical and non-informative complements?

- use frequency counts for dependency pairs
- normalize for lexical frequency
- $C * \log(p(a \to b)^2 / (p(a) * p(b)))$
- use thresholds for minimum co-occurence strength and minium absolute number of occurences
  - ask for strong positive correlation (mutual information)
  - $\log_2$ frequency classes: 1 (1), 2 (2-4), 3 (5-8), 4 (9-16) ....
- for a few special word classes, use generalizations:
  - PROP/hum (names)
  - NUM (numbers)
- separate treatment of pronouns (only relative to each other)

Data production



the applicational environment

- DeepDict is hosted at **www.gramtrans.com** where it is part of an integrated suite of translation tools
- it was developed as a spin-off from many years of
  - a) parser development
  - b) corpus annotation projects (Corpus Eye at corp.hum.sdu.dk)
  - c) lexicography and MT research
- The primary languages are the Germanic and Romance languages, but the method is largely language-independent given a lemmatized and dependency annotated corpus in that language
- Few similar approaches of turning corpus data into lexicography:
  - Sketch Engine Kilgariff et al. 2004)
  - Leipzig Wortschatz project (Biemann et al. 2004)

Corpus sources and parsers

	Parser	Lexicon	Grammar	Corpora
	DanGram	100.000 lexemes, 40.000 names	8.400 rules	ca. 159 M words (mixed)
	PALAVRAS	70.000 lexemes, 15.000 names	7.500 rules	ca. 210 M words (news) [+170 mill. wiki a.o.]
	HISPAL	73.000 lexemes	4.900 rules	ca. 90 M words (Wiki, Europarl, Internet)
	EngGram	81.000 val/sem	4.500 rules	ca. 210 M words (mixed) [+106 M email & chat]
	SweGram	65.000 val/sem	8.400 rules	ca. 60 M words (news, Europarl) [+ Wiki]
	NorGram	OBT / via DanGram	OBT / via DanGram	ca. 50 M words (Wikipedia) [+ internet]
	FrAG	57.000 lexemes	1.400 rules	-| [+67 mill Wiki, Europarl]
	GerGram	25.000 val/sem	LS+1.300 rules	ca. 44 M words (Wiki, Europarl) [+ internet]
	EspGram	30.000 lexemes	2.600 rules	ca. 58 M words (mixed)
	ItaGram	30.600 lexemes	1.600 rules	46 M (Wiki, Europarl)

cut-and-paste text

Languages: en, da, no, se, pt, eo

URL

Tools:
- WAP
- sms
- Firefox plugin
- Docs

File Edit View History Bookmarks Tools Help

http://gramtrans.com/

kabülisch

Gra... | GramTr... | yemeni... | My Acc... | GramTr... | Krak.dk | Welcom... | Aarhus ... | Norweg... | Result

Home News Products Languages DeepDict GramTrans

GRAMTRANS

Skip Navigation

Navigation

- Home
- News
- Products
  - Free
  - Personal
  - Commercial Lite
  - Commercial Standard
  - Commercial Direct
  - Schools
- Languages
  - System Features
  - Donate Words
- GramTrans
  - Contact
  - Feedback
  - Research

Tools

- Document Translation
- Mozilla Firefox Extension
- Texting (SMS)
- Statistics

Controls

- Site Admin
- My Profile
- Logout

News

- New lexicon feature
- New registration policy
- Substantial improvements for English - Danish

Text Translation

Danish to English | Clear | Translate | Length 0

Web Page Translation

http://

Danish to English | Translate

News

2008-04-11: New lexicon feature

Many people use GramTrans not just for machine translation, but also - or primarily - as a dictionary service, and for those users we have long provided the **QuickDict** feature - a simple lexicon lookup for individual words (including inflected words), where alternative translations can be preserved rather than contextually discarded. We have now gone one better and offer the **DeepDict** lexifier, a full "Learner's Dictionary" lookup providing usage information for the target language words, as well as example sentences and statistical information. Just click on the QuickDict translation (in red).

DeepDict is also available as such, of course, for direct lookup, and has its own page on the GramTrans site. For now, the QuickDict->DeepDict link works for the English side of Danish->English translations, while direct lookups also cover Danish and Spanish DeepDict, with more languages in the pipeline. As a target group we envision not just translators, but also schools, researchers and lexicographers. DeepDict is based on the grammatical and statistical analysis of huge text corpora, and can thus offer a unique and authentic overview of lexical usage, far beyond the scope of ordinary dictionaries.

2008-01-24: New registration policy

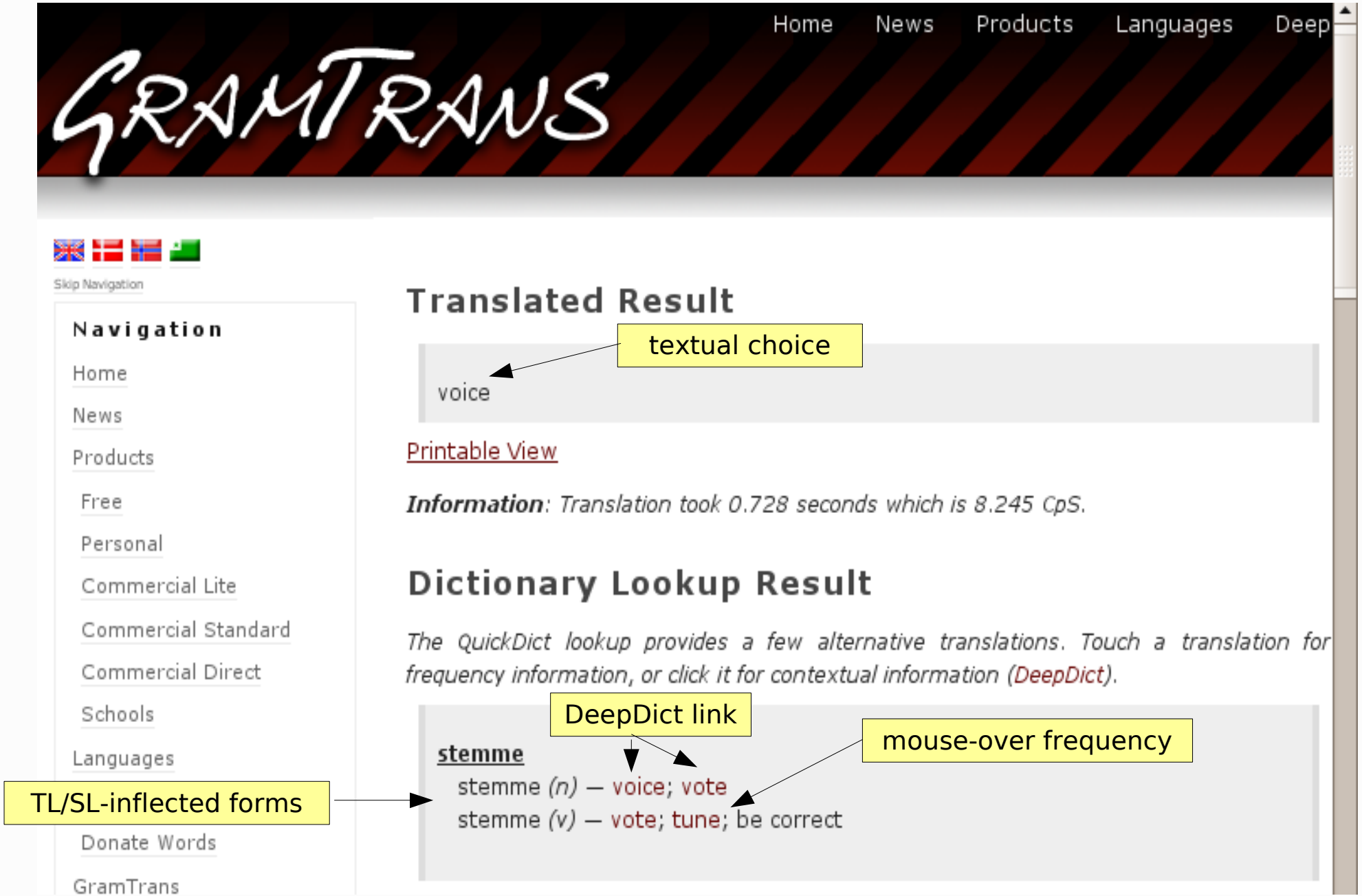
2008-01-03: Substantial improvements for English - Danish

2007-10-22: Testing Document Translation

Find: esperanto | Next | Previous | Highlight all | Match case | Reached end of page, continued from top

Done

The first dictionary layer: QuickDict



The second dictionary layer: DeepDict
Implicit semantics from lexical relations

Quick Reference

An entry consists of the following elements, taking "8.37:6 comic" as example:

- **8.37**: The co-occurrence strength between the lookup word and a given relation (relative frequency).
- **6**: Dual logarithmic value of absolute frequency. Scale is from 1 to 9.
- **comic**: The co-occurring word. In bold face if the absolute frequency value is 4 or higher.

Red numbers can be clicked to show examples in concordance form, if available.
For more information, please see the DeepDict Reference page.

Change Lookup Parameters

Word to look up:
vote

Word class:
- Noun
- Verb
- Adverb
- Adjective

Look up via DeepDict

Lookup language:
- Danish
- English
- French
- German
- Portuguese
- Spanish

Lexical frequency threshold:
- High
- Medium
- Low
- None

Minimum occurrence: 2

Minimum relative frequency: 0.0

Show top: 25

DeepDict: nouns 1

voice (noun)

countable

Premodifers:	PP postmodifiers:	Modifier of:
6.73:7 **loud** · 6.57:7 NUM · 6.41:6 **distinctive** · 5.05:7 **deep** · 6.64:5 **soprano** · 7.46:4 **gravelly** · 7.44:4 **husky** · 4.34:7 **single** · 5.21:6 **inner** · 4.2:7 **own** · 6.59:4 **baritone** · 5.58:5 **passive** · 6.52:4 **hoarse** · 4.46:6 **soft** · 5.46:5 **authoritative** · 4.32:6 **quiet** · 4.28:6 **human** · 6.28:4 **squeaky** · 6.16:4 **narrative** · 5.98:4 **gruff**	8.72:8 **rel-INDP** 1.35:3 interr-INDP 2.58:5 **of character** 2.43:5 **of reason** 2.25:5 **of god** 3.08:4 **from behind** 1.75:5 **of america** 2.7:4 **of conscience** 2.43:3 of dissent	6.04:7 **actor** · 5.94:4 **telephony** · 3.58:5 **actress** · 4.91:3 coil · 2.04:5 **communication** · 2.88:4 **talent** · 2.88:4 **recorder** · 3.52:3 choir · 2.19:4 **transmission** · 1.02:5 **vote** · 1.76:4 **channel** · 2.7:3 characterization · 3.59:2 synthesizer · 3.47:2 inflection · 3.41:2 synthesis · 0.31:5 **system** · 1.27:4 **message** · 1.6:3 directive · 0.51:4 **call** · 1.43:3 lesson

one can ...	14.93:2 modulate · 12.4:2 murmur · 8.62:5 **recognise** · 11.36:1 hush · 10.96:1 shriek · 3.44:8 **hear** · 9.28:2 amplify · 8.34:2 imitate · 3.72:6 **lower** · 6.7:3 obey · 7.2:2 mimic · 4.9:4 **lend** · 5.02:3 possess · 4.68:3 dub · 0.53:7 **raise** · 5.52:2 heed · 4.36:3 drown · 6.14:1 clip · 5.08:2 equal · 6.06:1 sharpen 8.54:4 **creep into** · 10.09:2 exclaim in · 9.09:2 mutter in · 2.63:8 **speak with** · 4.18:5 **sing in** · 8.07:1 retort in · 6.42:2 whisper in · 3.65:4 **reply in** · 4.24:3 cry in · 6.06:1 recite in · 5.78:1 startle by · 4.77:2 inject into · 2.77:4 **listen to** · 0.54:6 **speak in** · 3.39:3 detect in · 3.34:3 consist of · 2.91:3 shout in · 2.04:3 sing with · 0.78:3 sound like	**a voice**
a voice can ...	14.14:3 muffle · 12.44:4 **tremble** · 9.54:4 **whisper** · 11.44:2 crackle · 11.32:2 growl · 6.13:7 **sound** · 10.61:1 wobble · 9.49:2 drip · 9.49:2 thicken · 10.29:1 squeak · 7.85:3 falter · 5.82:5 **echo** · 8.69:2 waver · 7.43:3 harden · 8.24:2 reverberate · 8.7:1 exclaim · 5.38:4 **fade** · 5.25:4 **shout** · 6.17:3 deepen · 7.17:2 startle	
a voice can be	13.33:3 muffle · 12.39:2 hush · 8.1:3 clip · 9.75:1 tinge · 8.7:1 amplify · 1.58:7 **hear** · 4.68:3 dub · 5.12:2 choke · 4.09:2 drown · 3.62:2 strain	**...'ed**

DeepDict: nouns 2 - the “word field” side benefit

língua (noun)

countable

Premodifers:	PP postmodifiers:	Adjectival postmodifiers:
1.91:7 **próprio** · 0.63:6 **segundo** · 0.33:4 **só**	1.8:5 **de areia** 0.01:6 **de trabalho** 0.4:3 de difusão 0.12:1 de gringo	7.16:9 **português** · 6.6:9 **oficial** · 6.58:9 **inglês** · 3.85:8 **francês** · 5.79:6 **castelhano** · 3.91:7 **alemã** · 4.39:5 **gestual** · 5.17:4 **veicular** · 4.13:5 **nativo** · 2.71:6 **chinês** · 3.56:5 **latino** · 1.37:7 **nacional** · 2.17:6 **comum** · 2.12:6 **espanhol** · 2.01:6 **diferente** · 2.88:5 **albanês** · 3.76:4 **falado** · 3.57:4 **berbere** · 2.44:5 **curdo** · 1.29:6 **regional** · 1.99:5 **galego** · 1.74:5 **original** · 0.69:6 **local** · 2.47:4 **natal** · 2.44:4 **eslavo**

se pode ...	10.44:1 vivificar · 9.9:1 escovar · 2.79:7 **aprender** · 6.45:2 manejar · 6.44:2 afiar · -0.88:9 **falar** · 1.11:7 **dominar** · 4.88:3 morder · 4.64:3 desatar · 2.53:5 **ensinar** · 3.48:4 **soltar** 6.02:3 verter para · 1.53:7 **traduzir em** · 4.01:4 **redigir em** · 1.35:5 **exprimir em** · 0.17:6 **cantar em** · 0.09:5 **editar em** · 0.94:4 **imprimir em** · 0.78:4 **expressar em**	**uma língua**
uma língua pode ...	2.16:3 soltar · 0.46:4 **ensinar**	

Linked example concordances and “word sketches”

Forms	Abs Freq	Rel Freq
Total	6749	100.00%
vægt -> lægger	3608	53.46%
vægt -> lægge	1270	18.82%
vægt -> lagde	820	12.15%
vægt -> lagt	815	12.08%
vægten -> lægger	100	1.48%
vægten -> lægge	68	1.01%

Show all forms...

word sketch:

lægger	- stor mest afgørende meget	**vægt**	**på** {, / . / , fordi } {bag / på denne / på ved ansættelser / på det}

Concordances for: vægt_N -> lægge_V

ID	Text
inf30-33978	« Markedsføring på værdier og ansvar skal være en vigtig kilde til at øget konkurrenceevne på verdensmarkedet , fordi den globale forbruger i stadig større udstrækning **lægger vægt** på den slags , når han eller hun køber ind .
inf100-35624	I sidste ende giver den diplomatiske balancegang dog mulighed for , at den tyrkiske regering for en tid kan **lægge** mest **vægt** på sin egen tillægserklæring af indenrigspolitiske årsager , mens EU fokuserer på toldaftalen .
inf40-148330	« I dag **lægger** vi megen **vægt** på , at tillidsmanden skal arbejde med uddannelse og kunne svare på spørgsmål om pension .
c2000-dmfgjoab114	Netop den pind har vi fået ind , og det **lægger** vi stor **vægt** på , fordi vi netop vil gøre

collocation sketches 2

Form Statistics for: adoption_N -> give_V

Forms	Abs Freq	Rel Freq
Total	23	100.00%
adoption -> gave for	12	52.17%
adoption -> give for	7	30.43%
adoption -> gives for	3	13.04%
adoption -> given for	1	4.35%

word sketch:

	him up			
for	up	**for**	**adoption ->**	{. / . / , / and}
	up her baby			

Concordances for: adoption_N -> give_V

ID	Text
w7-1803406	The agencies will cover the costs of delivery and the medical care for any woman who **gives** up her baby **for adoption** .
w7-400518	Soon after release , she was impregnated by a man she barely knew and gave birth to a baby girl , which she had to **give** up **for adoption** .

DeepDict: Verb + Complements

caress (verb)

total of 527 relations

Subjects:	Accusative objects:
PERS: we, he, they, she 6.21:2 PROP · 4.79:2 finger · 4.62:1 breeze · 4.44:1 thumb · 2.89:2 hand · 1.47:1 eye	**PERS**: her, one another 6.62:2 cheek · 5.12:2 skin · 5.83:1 fingertip · 4.74:2 hair · 4.24:2 breast · 4.7:1 spine · 3.45:2 face · 4.42:1 jaw · 3.86:1 neck · 2.71:2 body · 3.71:1 PROP · 2.59:2 back · 3:1 length · 0.25:1 head

caress ...	5.54:2 gently · 3.71:1 sensuously
caress to ...	4.48:1 waist
caress with ...	4.01:1 tongue · 1.5:1 hand
caress in ...	0.23:1 way

Special treatment of word classes

- Each major PoS has its own lexicogram setup
  - verb + arguments/adjuncts, noun/adjective + modifiers
- PROP and NUM are generalized to avoid noise
- pronouns are very frequent and can't be directly compared to other lexical material
- but pronouns are carriers of abstracted semantic information (cp. Odense pronominal valency approach)
  - ± human: *who, what*
  - male / female: *he, she, him, her*
  - place, direction: *der, derhen, her, herhen (Danish)*
  - countable / mass: *much, many*

Pronouns as semantic classifiers

- “drikke” (drink): +anim vs. quantity
  - jeg (I), vi (we), han (he), ... den (UTR-it) <=> den (UTR-it), meget (much)
- “marry”: +hum/male vs. +hum/female, 2ps > 1ps
  - they, he, she, you, who, we <=> you, her, him, them, who, me
- “learn”: +hum <=> -hum/abstract
  - we, they, you, he, I, she <=> something, what, a lot, them, much
  - subclause complements: that-KS, interrogatives

learn (verb)

total of 63496 relations

Hide Frequencies

Subjects:	Subclauses:	Accusative objects:
PERS: we, they, you, he, I, she, i, who, one, one, everyone	4.95:9 **that**	**PERS**: many, something, what, a lot, them, much, anything, that,
9.92:9 **PROP** · 3.9:8 **child** · 4.72:5 **pupil** · 3.35:6 **student** ·	5.42:6 **interr**	nothing, all, which

Verb - adverb collocations

- (a) free adverb(ial)s: time, place, manner ...
- (b) valency bound adverb(ial)s
  - feel *how* (manner argument)
  - live *where* (place argument)
  - go *where* (direction argument)
- (c) verb-integrated particles
  - give *up,* fall *apart.* ? cut *out* (object predicative?)
- Since DeepDict is a lexicographical rather than a syntactic tool, we only keep verb-integrated particles separate (to allow sub-lemmatization), and lump everything else in an umbrella category (brown field)

“run” + adverbs

Verbal particles:
4.21:9 **out** · 4.62:8 **off** · 3.56:8 **down** · 2.68:7 **over** · 0.97:4 **through**

- known verbal particles

6.11:9 **away** · 5.64:7 **unsuccessfully** · 5.99:6 **counter** · 6.52:5 **aground** · 5.23:6 **midway** · 6.19:5 **concurrently** · 3.75:7 **fast** · 4.42:6 **homely** · 4.32:6 **north-south** · 4.68:5 **amok** · 4.61:5 **upstairs** · 2.51:7 **back** · 5.4:4 **northwards** · 3.32:6 **through** · 3.12:6 **south** · 3.06:6 **east** · 4.02:5 **smoothly** · 2.98:6 **west** · 1.94:7 **now** · 3.77:5 **east-west** · 4.76:4 **firstly** · 1.76:7 **well** · 2.57:6 **north** · 3.53:5 **quickly** · 1.51:7 **lately** · 1.03:7 **up** · 0.7:6 **down** · 0.29:5 **in**

- new verb-integrated particles: *run amok, run counter (to)*
- direction valency: *away, north-south, back, northwards*
- free manner adverbs: *unsuccessfully, quickly, smoothly*

mirrored by semantically distinct object complementation classes:

- *run (the) length / course (of) --- move adv.tr.*
- *run (n) miles --- move itr.*
- *run (the) risk --- fixed expression*
- *run (a) finger (along/over) --- move np.tr.*
- *run (a) program / system --- tr. “operate”*
- *run (a) school / centre / business -- tr. “organize”*

Verb - preposition collocates

- maybe the most ignored piece of usage information in dictionaries
- very difficult for learners, since the choice of preposition is more syntactic than semantic (cp. also aphasia research, Broca vs. Wernicke centres)
- like adverbs, prepositions (or rather pp's) can either be valency bound or free complements. It's near-impossible to make the distinction automatically, but know valencies are *-marked
- the binary dependency link has to be extended from the syntactic to the semantic head of the pp, storing 3-part links in the database

known valency-bound pp complements

run in ...	3.83:8 **election** · 3.93:5 **mode** · 2.73:6 **direction** · 0.19:3 median · 0.34:2 groove · 0.32:2 vein · 0.88:1
run at ...	6.47:5 **racecourse** · 0.11:4 **speed**
run *for ...	5.21:6 **re-election** · 4.81:6 **auditor** · 3.55:7 **gover** 3.27:6 **episode** · 1.14:8 **year** · 3.1:6 **commissione** 1.38:6 **leadership** · 1.33:6 **senate** · 1.32:6 **office** 1.49:3 governorship · 1.02:3 touchdown · 0.68:2 kilo
run *into ...	3.39:7 **trouble** · 3.58:6 **difficulty** · 1.33:6 **problem**
run *as ...	3.2:7 **candidate** · 1.09:3 independent
run at ...	3.53:6 **speed** · 1.43:4 **theatre** · 2.98:2 racecourse ·
run *on ...	3.25:6 **platform** · 1.83:5 **ticket** · 2.77:4 **petrol** · 0. 0.95:5 **track** · 1.94:4 **processor** · 0.7 0.63:3 mainframe · 0.6:3 architecture

free adverbial pp complements

run through ...	2.93:6 **hair** · 1.33:6 **term** · 1.03:6 **town** · 1.62 0.88:1 shire · 0.09:1 curl
run up ...	2.64:5 **stairs** · 0.56:3 steps · 0.19:1 overdraft
run over ...	2.1:5 **distance**
run along ...	1.69:5 **edge** · 0.28:5 **side** · 0.02:5 **line** · 0.58:
run under ...	1.35:5 **window** · 1.42:4 **banner** · 0.38:3 moti
run from ...	3.25:3 terminus · 1.67:4 **may** · 1.56:3 junction
run down ...	2.02:4 **stairs**
run on ...	0.96:5 **basis** · 0.48:1 microcomputer · 0.19:1 p

“drikke” (drink) + pp: implicit action frame

drikke med ...	3.13:5 **PROP-hum** · 2.07:1 svend · 0.
drikke i ...	3.02:4 **slurk** · 1.86:4 **PROP-top** · 0.1
drikke fra ...	1.04:5 **samling** · 0.35:4 **sans**
drikke til ...	0.08:4 **mad** · 0.46:1 aftensmÃ¥ltid
drikke af ...	2.07:1 plastickrus · 0.43:2 vandhane ·
drikke som ...	1.27:1 aperitif

social act of drinking together

manner: in sips,
where: plance names

fixed expression:
“drink s.o. unconscious”

drinking context: dinner

ritualized drink types:
starter drink

fra sig **fra** **sans ->** **og samling** {aftenen før brylluppet / og ryger til udpumpning til Vejle havn}
og {samling}

DeepDict: Verb + Prep.

vote ...	6.18:8 **therefore** · 4.87:8 **today** · 5.85:7 **tomorrow** · 4.24:6 **unanimously** · 3.62:5 **against** · 3.61:5 **overwhelmingly** · 1.49:7 **justly** · 2.31:5 **yesterday** · 4.22:3 firstly · 3.21:4 **accordingly** · 0.76:6 **now** · 3.47:3 tactically · 2.33:4 **differently** · 1.6:4 **separately** · 0.34:5 **however** · 0.16:5 **so** · 1.57:3 namely · 2.51:2 wholeheartedly · 1.37:3 actually · 0.26:4 **simple** · 0.93:3 freely · 0.89:3 naturally · 0.74:3 except · 1.53:2 intelligently · 0.43:3 narrowly
vote in ...	12.04:9 **favour** · 3.36:8 **election** · 3.66:6 **referendum** · 4.52:4 **entirety** · 1:6 **committee** · 0.61:5 **vote** · 0.07:4 **chamber** · 0.25:3 may · 0.06:1 plebiscite
vote against ...	5.36:9 **report** · 3.7:7 **resolution** · 3.66:7 **amendment** · 2.59:7 **proposal** · 3.58:6 **motion** · 3.03:4 **paragraph** · 2.33:4 **directive** · 1.6:4 **text** · 0.84:4 **recommendation** · 0.49:3 PROP-hum · 0.17:3 appointment · 1.16:2 accession · 0.53:2 discharge · 0.03:2 ratification
vote *for ...	5.06:9 **report** · 4.35:7 **resolution** · 3.67:7 **amendment** · 4.08:6 **motion** · 2.56:6 **reason** · 3.51:5 **directive** · 1.66:6 **candidate** · 3.22:4 **censure** · 1.12:5 **proposal** · 1.6:4 **text** · 2.24:3 accession · 0.54:4 **bush** · 0.03:4 **regulation** · 1.84:2 incorporation · 0.83:3 paragraph · 0.49:3 PROP-hum · 0.85:2 abolition · 0.51:2 postponement · 0.22:2 discharge · 0.01:2 tomorrow · 0.99:1 E · 0.82:1 deletion · 0.7:1 continuance · 0.7:1 guillotine · 0.5:1 assent
vote in ...	6.54:7 **favour** · 0.35:5 **poll**
vote *on ...	3.21:6 **amendment** · 1.97:7 **report** · 1.92:6 **resolution** · 2.38:5 **text** · 2.23:4 **directive** · 1.02:5 **proposal** · 1.01:5 **matter** · 1.34:4 **motion** · 1.72:3 paragraph · 0.53:4 **basis** · 1.42:3 enlargement · 0.29:3 statute · 0.25:3 may · 0.85:2 accession · 0.48:2 recital · 0.99:1 website · 0.01:1 scoreboard

structural symptoms of semantical differences

- meaning change between pre-modifying and post-modifying position (Romance languages)
- meaning change depending on head
  - *ill child/horse/relative* (state)
  - *ill omen/fate/fortune* (quality)
  - *ill wind/temper/humour* (intention?)
- adverbial premodifiers for adjective classification
  - intensity: *very*
  - measure: *<unit> noun moudifiers*
  - state: *temporal adverbs: often*
  - ±control: *intentionally*
  - result: *from, by, with*

DeepDict: Adjectives

pesado (adjective)

Pre-modifers:
10.18:9 **mais** · 2.25:7 **muito** · 2.05:6 **tão** · 1.54:6 **menos** · 2.32:5 **demasiado** · 2.61:4 **cada vez mais** · 1.02:5 **bastante** · 0.98:4 **algo** · 1.4:3 excessivamente · 0.33:4 **extremamente** · 0.83:2 um pouco · 0.16:2 de tal forma

Premodifier of:
5.13:7 **herança** · 5.07:7 **derrota** · 4.49:6 **multa** · 5.21:5 **fardo** · 3.06:6 **pena** · 2.92:5 **encargo** · 1.6:6 **responsabilidade** · 2.38:5 **sanção** · 3.23:4 **tributo** · 1.95:5 **carga** · 0.54:5 **estrutura** · 1.22:4 **indemnização** · 1.13:4 **perda** · 1.04:4 **condenação** · 0.98:4 **factura** · 1.91:3 sérvia · 1.51:3 ónus · 0.48:4 **silêncio** · 0.4:4 **tarefa** · 0.18:4 **custo** · 1.13:3 coima · 1.1:3 hum · 0.05:4 **dívida** · 0.64:3 bombardeamento · 0.57:3 burocracia

Postmodifier of:
7.34:8 **artilharia** · 7.1:8 **metal** · 6.3:8 **veículo** · 4.87:8 **arma** · 4.84:6 **armamento** · 3.38:6 **peso** · 4.17:5 **metralhadora** · 2.91:6 **pena** · 3.81:5 **comercial** · 2.77:6 **viatura** · 2.16:6 **estrutura** · 2.88:5 **herança** · 3.63:4 **maquinaria** · 2.45:5 **camião** · 2.24:5 **carga** · 2.2:5 **condutor** · 1.81:5 **derrota** · 1.76:5 **droga** · 1.22:5 **mão** · 1.9:4 **multa** · 0.88:5 **equipamento** · 1.59:4 **colisão** · 1.39:4 **motorista** · 0.39:5 **terreno** · 1.33:4 **castigo**

ill (adjective)

total of 4061 relations

Hide Frequencies

Pre-modifers:

9.03:7 **terminally** · 7.84:6 **seriously** · 6.11:6 **seriously** · 6.99:5 **gravely** · 4.37:7 **very** · 5.86:5 **critically** · 3.75:6 **too** · 4.67:5 **severely** · 4.8:4 **really** · 4.44:4 **physically** · 5.16:3 chronically · 4.14:4 **desperately** · 4.32:3 violently · 1.93:5 **so** · 2.76:4 **quite** · 3.47:3 dangerously · 3.02:3 critically · 2.92:3 dangerously · 2.92:3 extremely · 2.69:3 chronically · 3.46:2 acutely · 1.92:3 increasingly · 2.11:2 mortally · 1.07:3 extremely · 1.75:2 terribly

Post-modifiers:

4.23:4 **with fever**
2.66:4 **with cancer**
3.32:3 with tuberculosis
2.87:2 with pneumonia
3.42:1 with indigestion
3.42:1 with gout
1.99:2 with malaria
2.91:1 with cyst
1.75:2 with pleurisy

Post-modifiers:

4.23:4 **with fever**
2.66:4 **with cancer**
3.32:3 with tuberculosis
2.87:2 with pneumonia
3.42:1 with indigestion
3.42:1 with gout
1.99:2 with malaria
2.91:1 with cyst
1.75:2 with pleurisy

Premodifier of:

5.28:7 **health** · 7.86:3 repute · 4.7:6 **effect** · 4.67:6 **patient** · 5.4:4 **effects** · 5.91:3 omen · 3.37:5 **treatment** · 3.48:4 **fortune** · 5.33:2 humour · 2.55:3 luck · 2.49:3 feeling · 1.39:4 **person** · 0.59:4 **child** · 2.52:2 temper · 1.51:3 wind · 2.02:2 deed · 0.97:2 fame · 0.58:2 fate · 0.26:2 intention · 0.06:2 relative · 0.73:1 behoves · 0.53:1 judgement · 0.2:1 fit · 0.11:1 reception · 0.08:1 grace

Semantically motivated collocation restrictions: usage of *big / high / large*

high ...	**big ...**	**large ...**
• level • [school] • concentration • speed • proportion • altitude • elevation • temperature	• [bang, band] • hit • problematic • break • difference • brother • star, bird • man, city	• number • quantity • amount • proportion • sum • portion, part • city, island • population
<degree> <measure>	<size> <importance>	<extension> <quantity-mass>

direct collocates vs. dependency collocates

- easy to catch: multi-word expressions (MWE), or what in English amounts to “compounds”, often with phonetic stress on the first word
  - big bang, big band
  - high tide, high society
- more difficult, often non-adjacent, profits from dependency relations
  - high temperature
    - **high** room **temperature**
    - **ambient** temperature was rather **high** when ...

Bilingual polysemy / equivalence check: “caress” objects in Danish / Swedish

- DANISH: kærtegne
  - **body parts**: bryst, krop, kind, hud, balder, mave, inderlår, brystvorte, hår, ansigt, klitoris, lår, sexbombe, nosse, røvhul, nakke, hals, kropsdel, bagdel
  - **surfacees:** silkestof, græsbane
  - **PROP**-hum

DAN: stryge

* swipe
* iron
* remove
* move ...

- SWEDISH: smeka
  - **body parts**: kind, kønsorgan, brøst, stjärt, klitoris, kropp
  - **PROP**-hum
  - **ball:** boll, passning, tennisboll
  - **instrument**: elgitarr
  - **things:** lack, rännil, instrmentpanel, julle, murbrok, vidunder

SWE: stryka

DeepDict: Verb + semantic classes (only compiled for Danish, Norwegian and Esperanto)

- uses ~ 200 semantic prototype classes for nouns
- correlations are computed the same way as for words
- offers a level of abstraction, and can compensate for sparse data

drikke (verb)

total of 41235 relations

Hide Frequencies

Subjects:

PERS: jeg, vi, han, de, man, du, der, hun, den

7.48:8 **<H>** · 7.38:8 **PROP-hum** · 5.9:7 **<Hprof>** · 5.6:6 **<Hnat>** · 5.4:6 **<Hfam>** · 5.38:6 **<HH>** · 5.48:3 PROP · 1.25:5 **dansker** · 1.74:4 **<Hideo>** · 1.62:4 **<Azo>** · 1.62:4 **<A>** · 0.12:5 **ung** · 1.21:3 PROP-org · 2.07:2 alkoholafhængig · 1.05:3 <Adom> · 0.68:3 <Lh> · 0.24:3 <food> · 1.37:1 medhustru · 1.37:1 nescafé · 0.39:1 Forældrene

Accusative objects:

PERS: den, meget

11.48:9 **<drink-h>** · 9.83:9 **<drink-m-h>** · 9.2:9 **<drink-c-h>** · 8.88:9 **<cm-liq>** · 8.78:9 **<occ>** · 7.24:9 **kaffe** · 7.47:8 **<cm-liq-h>** · 7.2:8 **<B>** · 6.93:8 **øl-2** · 8.8:5 **PROP** · 5.73:8 **vin** · 6.51:7 **øl-1** · 6.31:7 **<amount>** · 6.24:7 **<drink>** · 6.23:7 **te** · 5.52:7 **alkohol** · 6.13:6 **bajer** · 4.05:8 **vand** · 5.44:6 **<con>** · 5.29:6 **PROP-hum** · 5.28:6 **cola** · 5.05:6 **<cc-h>** · 4.91:6 **<ac>** · 4.87:6 **rødvin** · 4.68:6 **whisky** · 4.46:6 **<drink-m>** · 4.16:5 **<mat-h>** · 4.13:5 **<cc>** · 3.9:5 **<drink-c>** · 3.73:5 **PROP-org** · 3.45:5 **<H>** · 3.4:5 **<HH>** · 3.4:5 **<food-h>** · 3.19:5 **<am>** · 3.01:5 **PROP-top** · 3.01:5 **<Hprof>** · 2.67:4 **<temp>** · 2.59:4 **<cm-h>** · 2.26:4 **<food-m-h>** · 2.17:4 **<cm>** · 2.17:4 **<unit>** · 1.86:4 **<anorg>** · 1.62:4 **<per>** · 1.05:3 <Hfam> · 1.05:3 <f-psych>

Verbal particles:

4.92:6 **ud** · 3.03:6 **ihjel** · 1.28:6 **sammen** · 0.45:5 **op** · 1.47:3 ned · 3.41:1 bort

“drikke” (drink) - object classes

- Words:
  - kaffe
  - vin
  - øl
  - te
  - alkhol
  - bajer
  - vand
  - cola
  - rødvin
  - whisky

- Semantic prototypes
  - <drink-h>, <drink-m-h>, <drink-c-h>, <drink>, <drink-m>, <drink-c>
  - <cm-liq> *(vand, urin)*
  - <amount> *(masse, mundfuld)*
  - <con> *(kop, flaske),* <unit> *(liter)*
- ambiguity artifacts:
  - kaffe <occ> <Lh>, glas <mat-h>
- metaphor:
  - <anorg> (“drink one's brain out”)
- cross-class:
  - <food-h> *(gift)*

The background: 10 dependency parsers

Language	Parser	Lexicon	Analyzer	Grammar	Levels	Applications
	DanGram	100.000 lexemes, 40.000 names	Full	8.400 rules	morph., syntax, dep., psg, sem. roles	Teaching, corpus annotation, MT, Spell/Grammar checker, QA-systems, NER
	PALAVRAS	70.000 lexemes, 15.000 names	Full	7.500 rules	morph., syntax, dep., psg, sem. roles	Teaching, corpus annotation, MT, QA-systems, NER
	HISPAL	73.000 lexemes	Full	4.900 rules	morph., syntax, dep., psg, sem. roles	Teaching, corpus annotation
	EngGram	81.000 val/sem	Full	4.500 rules	morph./syntax, dep., psg	Teaching, corpus annotation, MT
	SweGram	65.000 val/sem	Full	8.400 rules	morph./syntax, dep., psg	Teaching, corpus annotation, MT
	NorGram	OBT / via DanGram	Full	OBT / via DanGram	morph./syntax, dep., psg	Teaching, corpus annotation, MT
	FrAG	57.000 lexemes	DTT + analysis	1.400 rules	morph.-correction, syntax, dep., psg	Teaching, corpus annotation
	GerGram	25.000 val/sem	Full (Lingsoft)	LS+1.300 rules	morph. (Lingsoft), syntax, dep., psg	Teaching, corpus annotation
	EspGram	30.000 lexemes	Full	2.600 rules	morph., syntax, dep.	Teaching, corpus annotation, MT
	ItaGram	30.600 lexemes	DTT + analysis	1.600 rules	morph., syntax, dep.	Teaching, corpus annotation

How to use DeepDict 1

- as a lexicographer
  - find inspiration as to **complementation patterns** (selection restrictions) for dictionary entries
  - find the **most typical** (not just the most common!) **example** of a certain construction
  - find candidates for **multi-word expressions**
  - find candidates for **metaphorical usage** (often high correlation index because one of the parts is infrequent on its own)
  - find **semantic distinctions** and **subsenses** not otherwise obvious, and triggered by head or dependent words (e.g. mistænksom – mistænkelig, high – big - large)

How to use DeepDict 2

- for teaching
  - create **lexical fields**
    - e.g. edibles, drinkables etc., via 'eat', 'drink')
    - a list of languages? countries? professions?
    - all about language
    - horse/share/oil-related words for an essay: what does it do (SUBJ), what do you do with it (ACC), how is it characterized (prenominals)
  - find **phrasal verbs** / **prepositional complements**
  - describe **usage differences** between near synonyms
  - distinguish between **literal** and **abstract** uses (e.g. heavy – tung – schwer) ... are some of these cross-language phenomena?
  - find **metaphors** (caress_V)
  - find **gender differences** through pronouns (caress_V)

Sociolinguistic exercise: *my new neighbour is a ...* refugee - fugitive - immigrant - foreigner

- **immigrant:** illegal, German, Irish, Italian, Jewish, NUM, Chinese, recent, European, Polish, legal, undocumented, Mexican
  - subculture, (il)legality
- **foreigner:** NUM, unauthorized, undesirable, untidy, barbarian, meddlesom, friendless, unreliable, stateless, naturalized
  - (negative) focus on (lack of) assimilation
- **refugee:** Afghan, Palestinian, Jewish, genuine, Vietnamese, Kurdish, Politcal, protestant Albanian, NUM, huguenot, Palestinian, Rwandan
  - recent, reason focus
- **fugitive:** wanted, al-Qaeda, ephemeral, NUM, hunted, highest-ranking, royalist, exhausted, Russian, harried, displaced
  - process/context focus (war, factions, persecution)

Perspectives 1: Lexicography

- DeepDict shows how syntactically related word pairs can be “harvested” from dependency- and function-annotated corpora
- It allows the lexicographer
  - not only to **find examples and frequencies** for certain (known) collocations and lexical constructions, but also
  - to **compile *new* lists** of such collocations and constructions
- DeepDict is a language-independent method, and could be built not only for further languages, but also for specialized or customer-built corpora (genre-variation, diachronic variation, spoken language, specific author)

Perspectives 2: Grammars

- Better parsers, specifically, better parser lexica
- DeepDict databases can be used to harvest
  - **valency** patterns
  - semantical **selection restrictions** (generalized from words or prototypes)
  - likelihood thresholds for semantical fillers of syntactic slot
- Cyclical interplay between DeepDict-style corpus information and the parser(s) that provided it

e.g. Portuguese lexicon entry:

- pensar ('think'): <fSUBJ/H:74>, <FSUBJ/org:25>

used in grammars, e.g. +hum marking in anaphor grammar:

- ADD (£hum) TARGET PERS + @P<
  (p @PIV LINK 0 PRP-COM LINK p (<fPRP-com/H>70>))

Perspectives 3: FrameNet

- current DeepDict: shows one relation at a time, i.e. computes e.g. subject and object fields independently of each other, which is fine for many applications, but could be taken a step further using:
- lexico-semantic frames (Berkeley FrameNet)
  - @SUBJ / <hum> ==> “**read**” <vt> ==> @ACC / <sem-r>
- same corpus annotation needs as for DeepDict
  - verbal subsenses need to be specified and semantically classified
  - senses can be structurally corroborated or supplemented interactively from corpus data
- project site: www.framenet.dk

DeepDict
@
gramtrans.com

CorpusEye: corp.hum.sdu.dk
Parsers: visl.sdu.dk

eckhard.bick@mail.dk

Spin-offs: WebPainter

- live in-line markup of web pages
- mouse-over translations while reading

File Edit View History Bookmarks Tools Help

ike_race&pair=eng2dan&x-form-id=webpaint_url | charmed

Tr... | My Acc... | Welcom... | http... | Norweg... | Bicycle ... | Google ... | GramTr...

GramTrans WebPainter | http://en.wikipedia.org/ | English to Danish | WebPaint | Return to front page

mouse-over translation:

trekanter (trekant)

All of the above bicycle races involve diamond frame bicycles of two triangles . An alternative is the recumbent , a bicycle on which the rider sits back with the legs horizontal . This puts the body in a position where there is less wind drag . Proponents claim it provides more ortable riding , with no weight on the wrists . The recumbent is a more aerodynamic design of bicycle , and world speed records were set with them .

Bicycle races are popular all over the world , especially in Europe . The most devoted countries are Italy , Spain , Belgium , Germany , France , the Netherlands and Switzerland , although the United States has international standing , as does Australia . The USA boasts three-time Tour de France and first American winner , Greg LeMond as well as seven-time winner Lance Armstrong . Australia has

In Europe bicycle racing expresses nation prestige : German Democratic Republic postage stamp depicting Gustav Adolf Schur (Täve) , 1960

optional grammar (here: SUBJ and **prep**