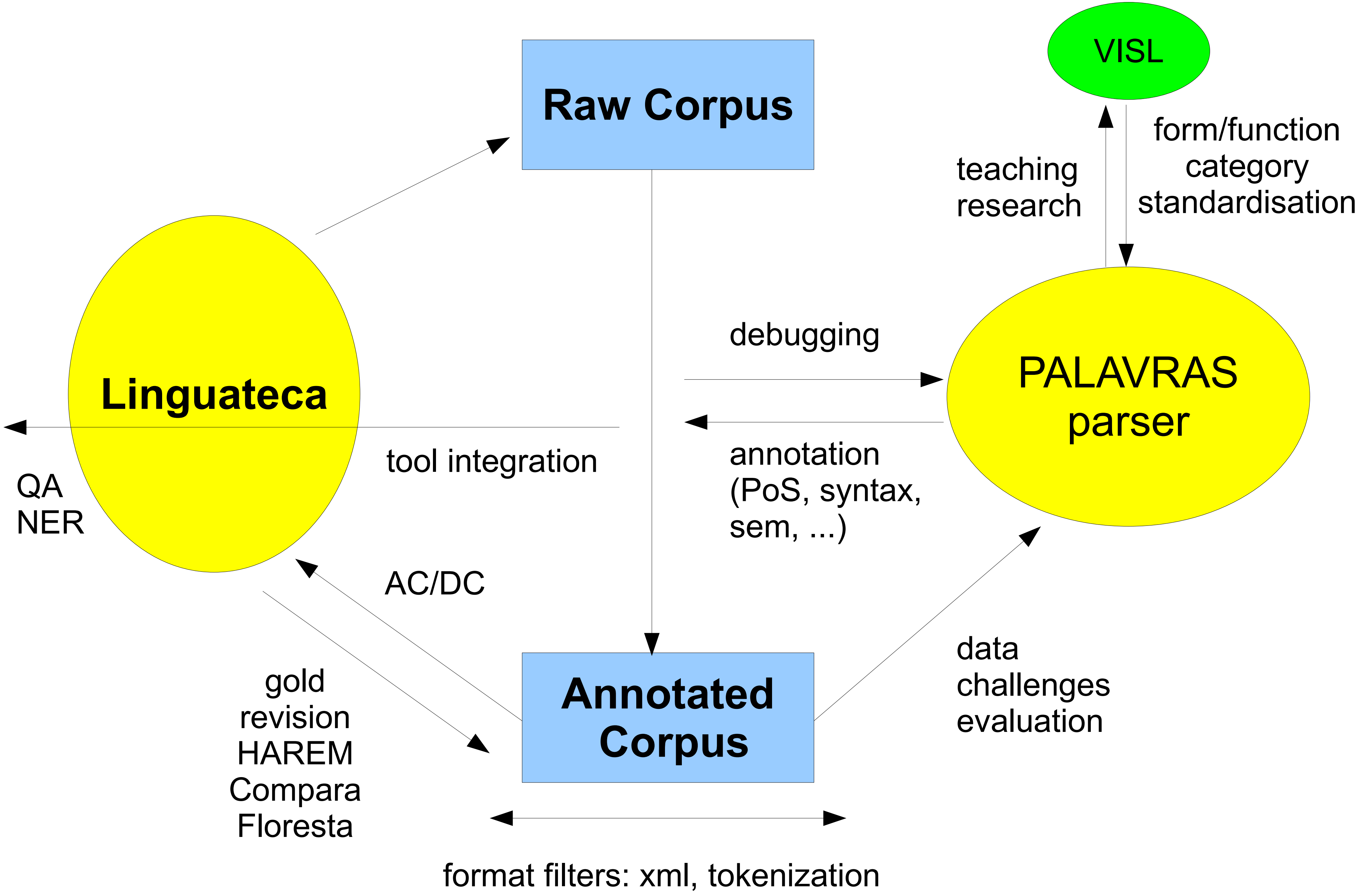
The PALAVRAS parser and its Linguateca applications - *a mutually productive relationship*

Eckhard Bick
University of Southern Denmark
eckhard.bick@mail.dk

Outline

- Flow chart Linguateca <-> Palavras
- History and milestones
  - pre- and co-Linguateca
- Parallel development strands
- A recent example of ressource integration: QuickDict
  Público + Folha + Dependency Parsing + Statistics = Lexicography



Raw Corpus

VISL

teaching
research

form/function
category
standardisation

Linguateca

debugging

PALAVRAS
parser

tool integration

annotation
(PoS, syntax,
sem, ...)

QA
NER

AC/DC

data
challenges
evaluation

Annotated
Corpus

gold
revision
HAREM
Compara
Floresta

format filters: xml, tokenization

History of the parser (pre-Linguateca)

- 1990-93 Portuguese lexicography (master's thesis)
- 1992-93 Morphological analyzer for Portuguese
- 1994-96 Development of the first version of PALAVRAS' Constraint Grammar modules (morphology and syntax)
- 1996 presentation at the 2. Propor in Curitiba, Brazil
- 1997 focus on tree structures and corpus work (Borba-Ramsey, VEJA, NILC)
- 1998-99 experiments with heuristics, transscribed speech (NURC), dialectal variation (Cordial-syn, Moçambique) and historical data (Tycho Brahe)
- 1997-? PALAVRAS and derived systems for other languages used for syntax teaching in Denmark (VISL)
- 2000 Dr.phil. thesis centering on PALAVRAS

History of the parser (co-Linguateca) - corpus annotation

- 1999-? AC/DC project (e.g. Santos & Bick 2000, LREC)
  - both Portuguese (esp. Público) and Brazilian (Folha) data
  - quantitatively mostly newspaper genre
  - mostly unrevised annotation, but feedback and multiple reannotations
- 2000-? Floresta Sintá(c)tica Project (e.g. Afonso et al. 2002, LREC)
  - 3 phases: (a) Odense-Oslo, (b) multipolar international, (c) Portugal
  - automatic annotation with human linguistic revision
  - (a-b) PALAVRAS + PSG (Bick 2003, CL), two-pass annotation and revision
  - (c) PALAVRAS + DG (Bick 2005, TLT) , one-pass revision
- 2004 PALAVRAS-Linguateca license at SINTEF, Oslo
  - raw and edited annotation (e.g. COMPARA)

History of the parser (co-Linguateca) - joint evaluation

- 2003 Morfolimpíadas/Avalon (Santos 2007)
  - PALMORF, an adapted morphology module from PALAVRAS, achieves top results
- 2005 1. HAREM (evaluation meeting 2006, Porto)
  - PALAVRAS-NER (Bick 2006, Propor Itatiaia), with separate modules for name recognition and classification, participates with good results (best F-scores)
- 2008 2. HAREM
  - the SeReLep system (by PUCRS) integrates a licensed PALAVRAS

Other development strands

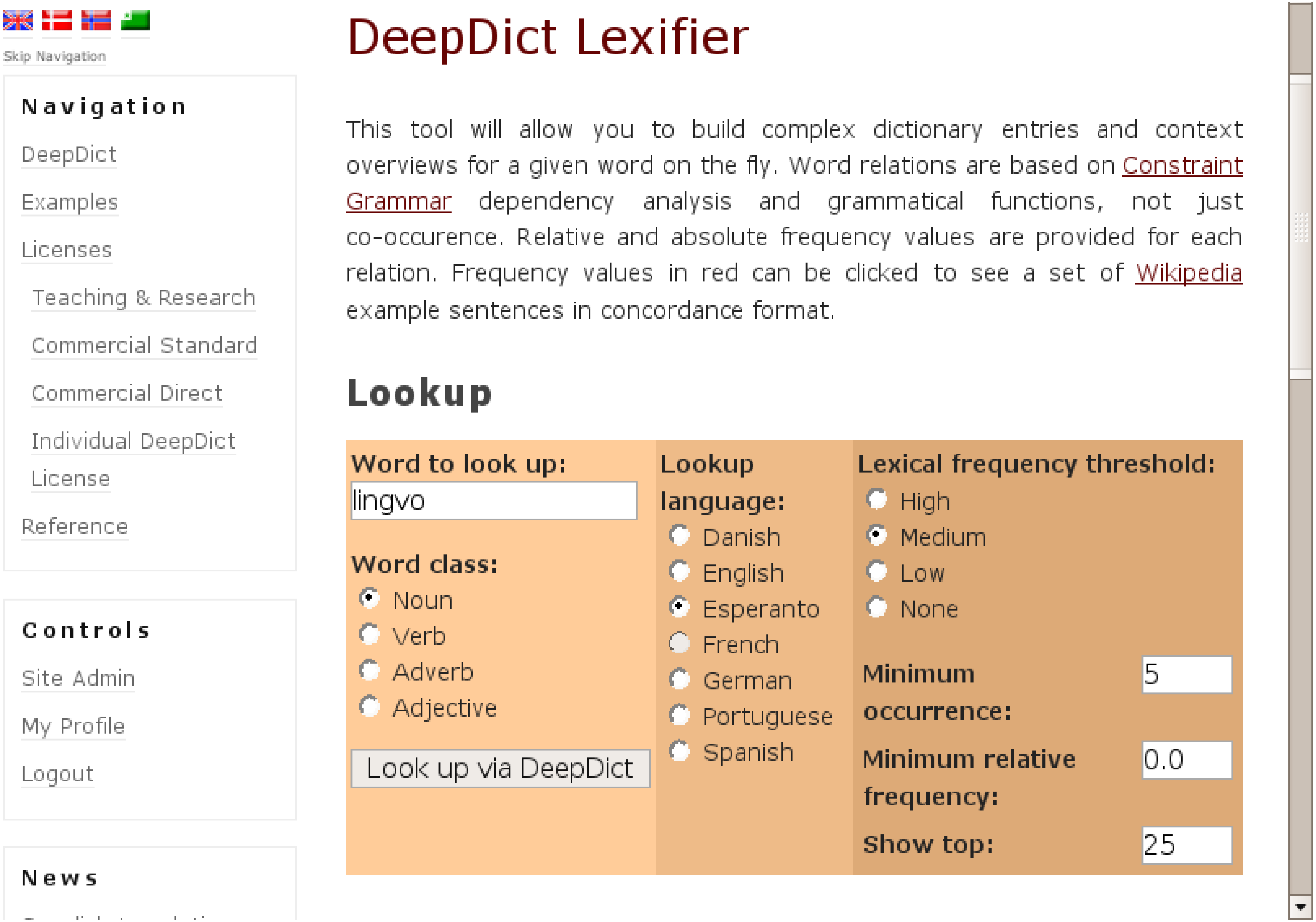
- Question & Answering
  - A prototype using PALAVRAS syntactic analysis (Bick 2003, EPIA)
  - PALAVRAS used as a syntactic module in other systems
    - University of Evora (Quaresma et al. 2004, CLEF)
    - Esfinge (Costa 2006, CLEF)
- Historical Portuguese
  - Parser and lexicon daptations for 18th and 19th century Brazilian Portuguese (Bick & Módolo 2005)
- Semantic annotation beyond NER
  - Semantic prototype annotation (used for MT and Floresta)
  - Semantic role annotation (Bick 2007, TIL)
- CG3, a new Constraint Grammar formalism and compiler
  - direct creation and referencing of dependency and anaphora relations, context windows larger than a sentece, reg. expressions, integration of statistical data, ...

DeepDict

- A lexicographic tool to provide contextual dictionary information on the fly – useful for dictionary publishers, students, linguists ..
- example of a product integrating different types of ressources into a real life tool: Corpus data, Tagger/Parser, Statistical tools
- example of interactive, corpus-driven lexicography
  - (1) results can be fed back into the parsing lexicon (valency tags, semantic lumping etc)
  - (2) the improved parsing lexicon allows better corpus annotation and – in turn – more DeepDict data
- the Portuguese version is freely available on the Internet

DeepDict
www.gramtrans.com

- syntactically analyzed corpus (dependency links and functions)
  - Linguateca's Público and Folha corpora (CETEMPúblico, CETENFolha) – ca. 180+30 M words
  - Portuguese Wikipedia (Nov. 2005) – ca. 8 M words
  - Portuguese section of Europarl – ca. 27 M words
- lemmatization, “normalization” (passives, numbers, names)
- extraction of mother-daughter relations, **depgrams**, not ngrams (cf. Adam Kilgariff's Sketch Engine)
  - N + @N< (vaca *louca*), @>A + ADJ (*gravemente* doente)
  - V + @ACC (ganhar *terreno)*, @SUBJ ktp, V + PRP (pensar *em)*
- co-occurrence measure: p(AB) / p(A) * p (B)
- graphical interface



DeepDict Lexifier

This tool will allow you to build complex dictionary entries and context overviews for a given word on the fly. Word relations are based on Constraint Grammar dependency analysis and grammatical functions, not just co-occurence. Relative and absolute frequency values are provided for each relation. Frequency values in red can be clicked to see a set of Wikipedia example sentences in concordance format.

Lookup

Word to look up: lingvo

Word class:
● Noun
○ Verb
○ Adverb
○ Adjective

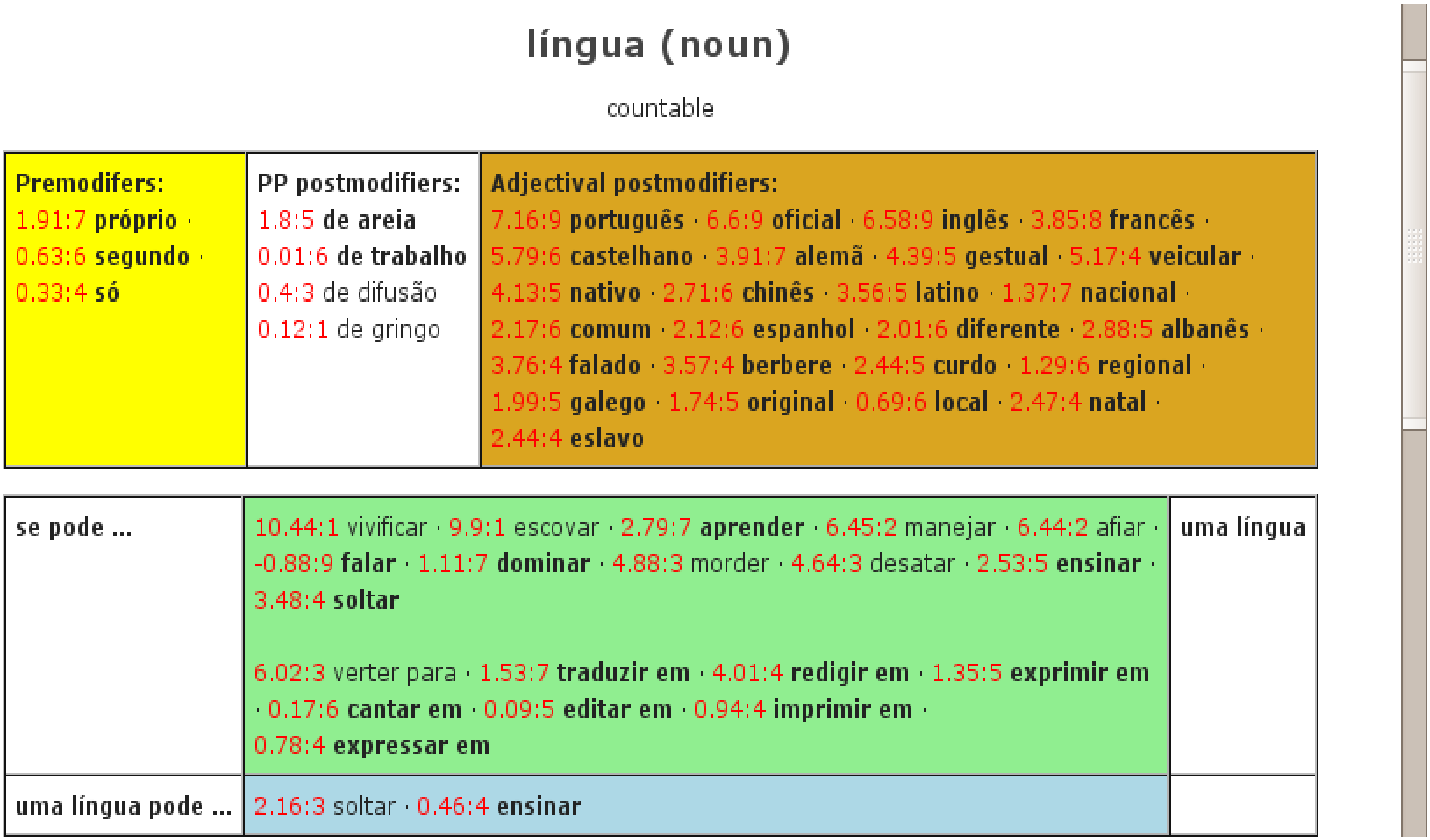
Lookup language:
○ Danish
○ English
● Esperanto
○ French
○ German
○ Portuguese
○ Spanish

Lexical frequency threshold:
○ High
● Medium
○ Low
○ None

Minimum occurrence: 5

Minimum relative frequency: 0.0

Show top: 25

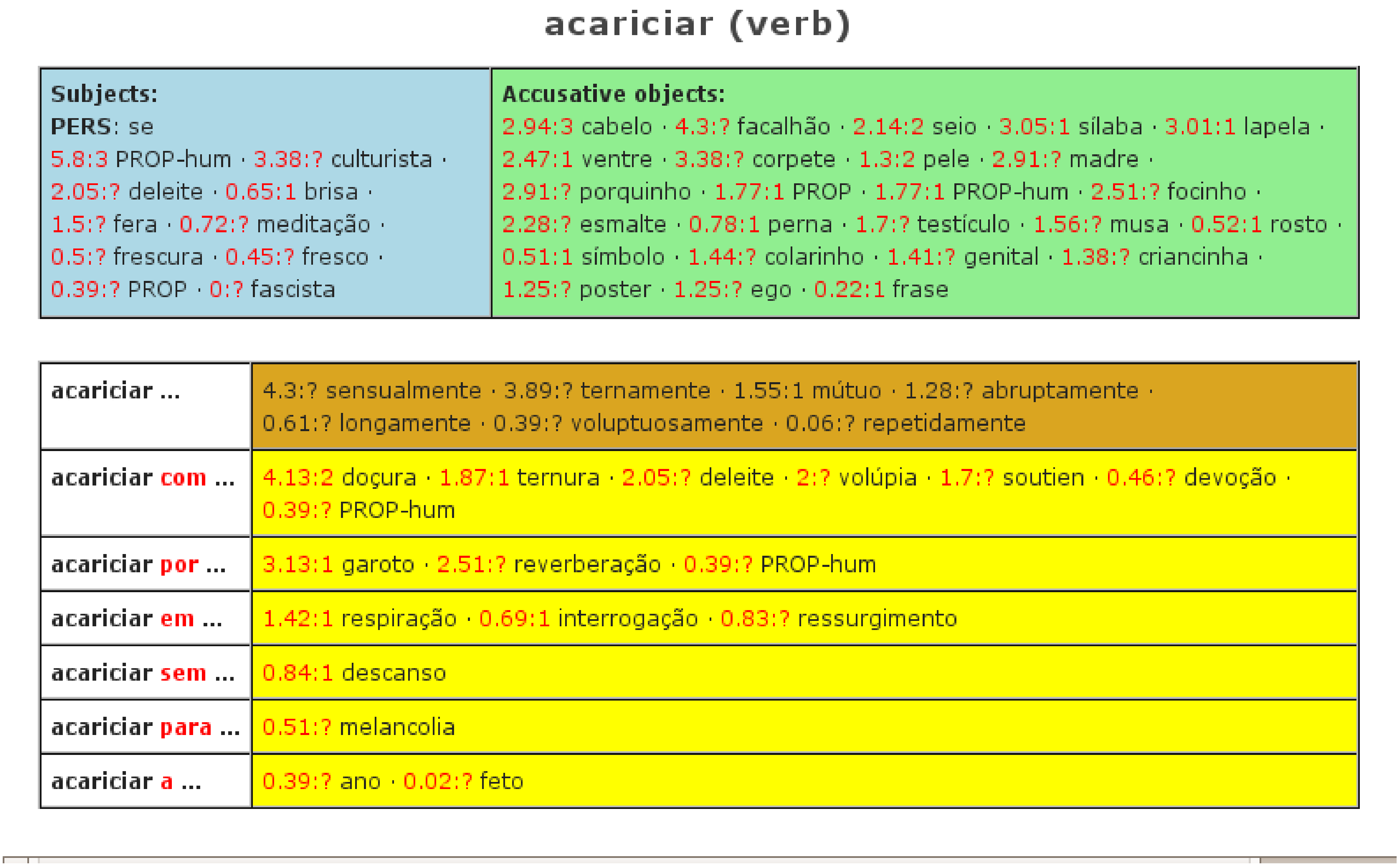


língua (noun)

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**	

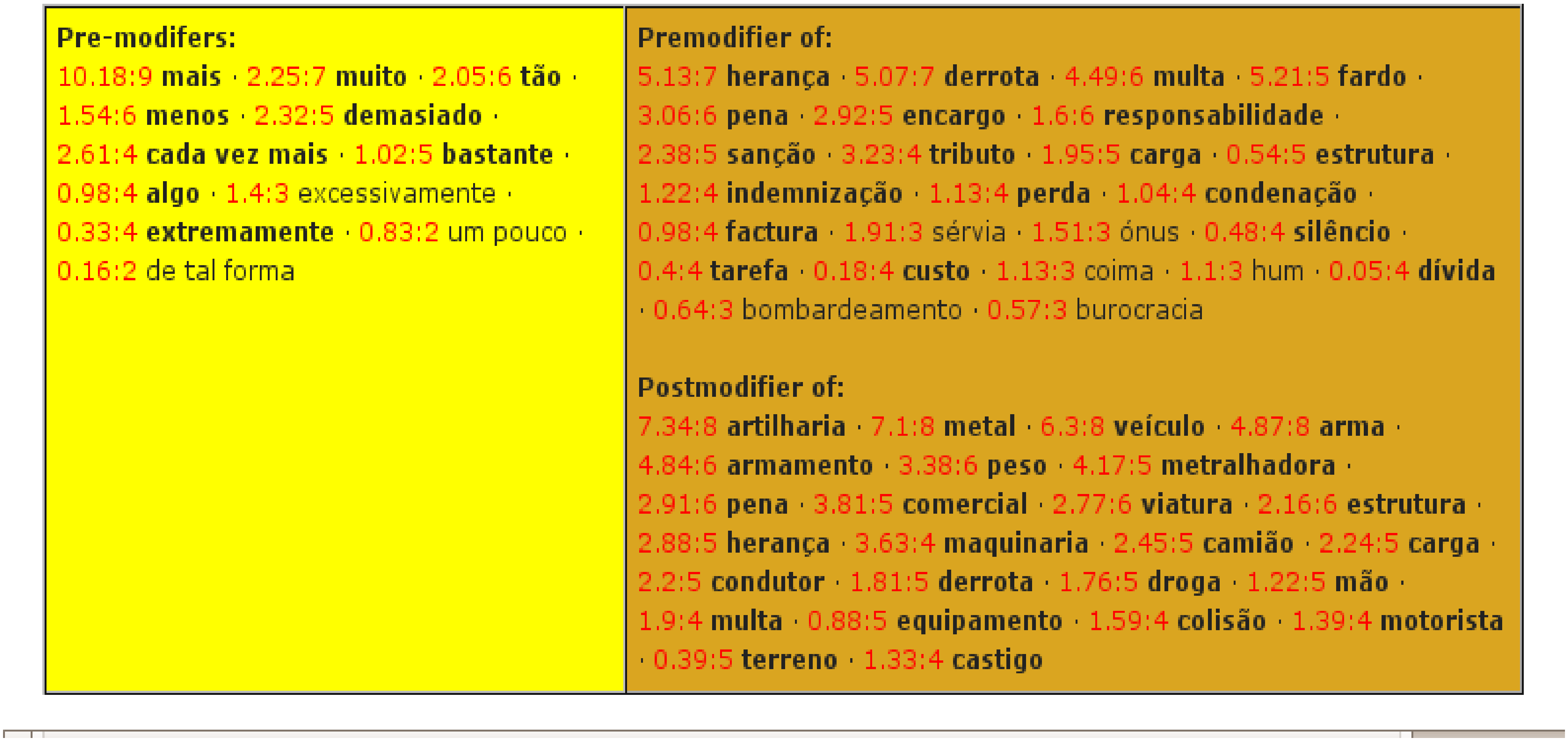


acariciar (verb)

Subjects:	**Accusative objects:**
PERS: se 5.8:3 PROP-hum · 3.38:? culturista · 2.05:? deleite · 0.65:1 brisa · 1.5:? fera · 0.72:? meditação · 0.5:? frescura · 0.45:? fresco · 0.39:? PROP · 0:? fascista	2.94:3 cabelo · 4.3:? facalhão · 2.14:2 seio · 3.05:1 sílaba · 3.01:1 lapela · 2.47:1 ventre · 3.38:? corpete · 1.3:2 pele · 2.91:? madre · 2.91:? porquinho · 1.77:1 PROP · 1.77:1 PROP-hum · 2.51:? focinho · 2.28:? esmalte · 0.78:1 perna · 1.7:? testículo · 1.56:? musa · 0.52:1 rosto · 0.51:1 símbolo · 1.44:? colarinho · 1.41:? genital · 1.38:? criancinha · 1.25:? poster · 1.25:? ego · 0.22:1 frase

acariciar ...	4.3:? sensualmente · 3.89:? ternamente · 1.55:1 mútuo · 1.28:? abruptamente · 0.61:? longamente · 0.39:? voluptuosamente · 0.06:? repetidamente
acariciar com ...	4.13:2 doçura · 1.87:1 ternura · 2.05:? deleite · 2:? volúpia · 1.7:? soutien · 0.46:? devoção · 0.39:? PROP-hum
acariciar por ...	3.13:1 garoto · 2.51:? reverberação · 0.39:? PROP-hum
acariciar em ...	1.42:1 respiração · 0.69:1 interrogação · 0.83:? ressurgimento
acariciar sem ...	0.84:1 descanso
acariciar para ...	0.51:? melancolia
acariciar a ...	0.39:? ano · 0.02:? feto

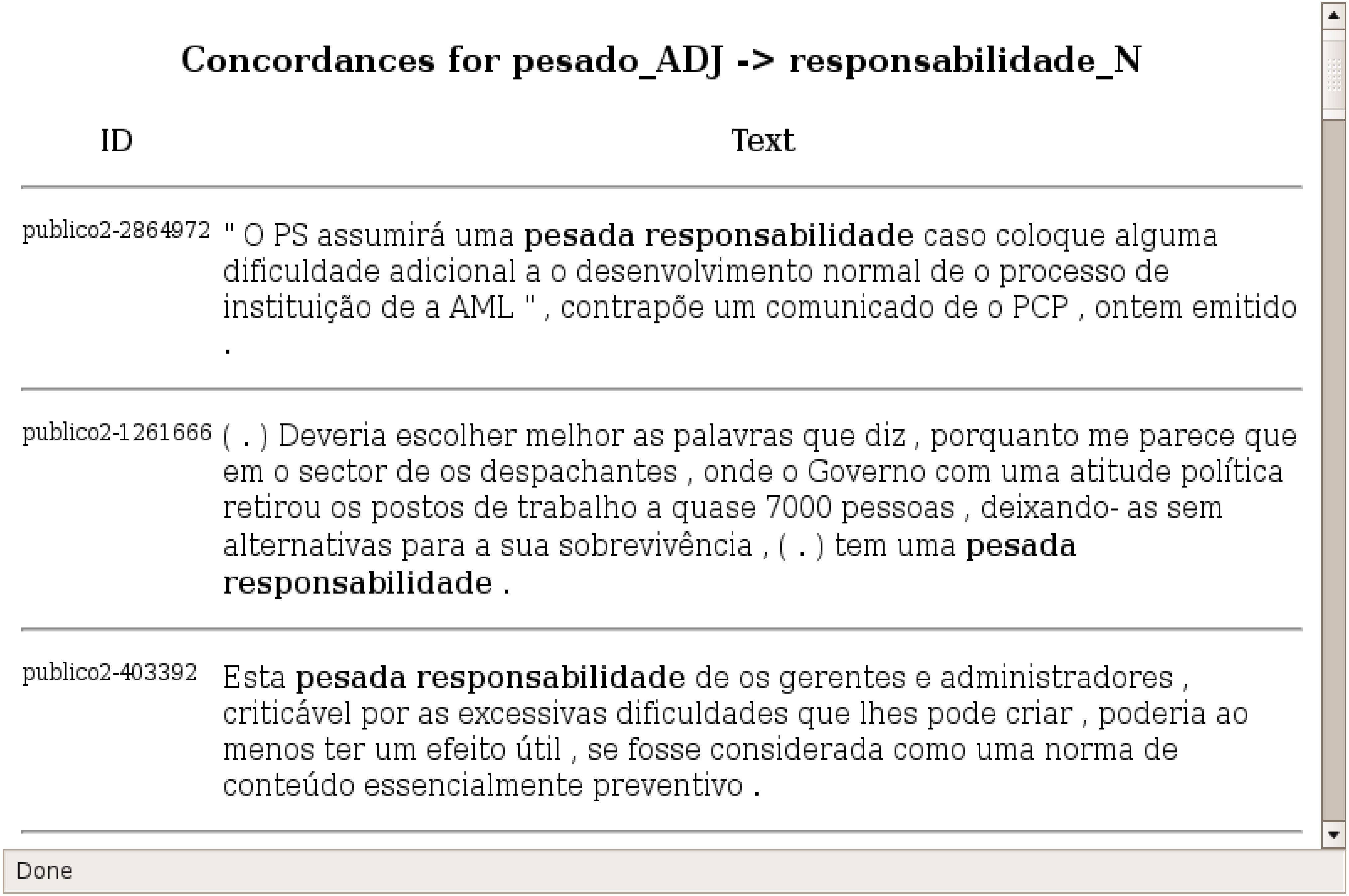
pesado (adjective)



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**



Concordances for pesado_ADJ -> responsabilidade_N

ID	Text
publico2-2864972	" O PS assumirá uma **pesada responsabilidade** caso coloque alguma dificuldade adicional a o desenvolvimento normal de o processo de instituição de a AML " , contrapõe um comunicado de o PCP , ontem emitido .
publico2-1261666	(.) Deveria escolher melhor as palavras que diz , porquanto me parece que em o sector de os despachantes , onde o Governo com uma atitude política retirou os postos de trabalho a quase 7000 pessoas , deixando- as sem alternativas para a sua sobrevivência , (.) tem uma **pesada responsabilidade** .
publico2-403392	Esta **pesada responsabilidade** de os gerentes e administradores , criticável por as excessivas dificuldades que lhes pode criar , poderia ao menos ter um efeito útil , se fosse considerada como uma norma de conteúdo essencialmente preventivo .

- http://beta.visl.sdu.dk/
  - visl/pt/parsing/automatic/ *(live parsing, file upload etc.)*
  - constraint_grammar.html *(formalism)*
  - visl/pt/info/ *(categories and annotation docs)*