# Words and their secrets

**Diana Santos & Maria José Bocorny Finatto**

**ESSLLI 2010, Copenhagen**

# References for "Words and Their Secrets" at ESSLLI 2010

Diana Santos & Maria José Bocorny Finatto

**Works cited in the course**

Baayen, R. Harald. *Word frequency distributions*. Kluwer Academic Publishers, 2001.

Bacelar do Nascimento, Maria Fernanda, José Bettencourt Gonçalves, Lucília Chacoto, Paula Neto & Luísa Alice Santos Pereira. "Ambiguidade morfológica no Português Fundamental". In *Actas do 1º Encontro de Processamento da Língua Portuguesa (escrita e falada) (EPLP'93)* (Lisboa, 25-26 de Fevereiro de 1993), 1993, pp. 101-106.

Bick, Eckhard. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

Bindi, Remo, Nicoletta Calzolari, Monica Monachini, Vito Pirrelli & Antonio Zampolli. "Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition", *Literary and Linguistic Computing* **9**, No. 1, 1994, pp. 29-46.

Bortolini, U., C. Tagliavini & A. Zampolli. *Lessico di frequenza della lingua italiana contemporanea*. IBM Italia, 1981.

Bowerman, Melissa. "The origins of children's spatial semantic categories: cognitive versus linguistic determinants", in John Gumperz & Stephen C. Levinson (eds.), *Rethinking linguistic relativity*. Cambridge University Press, Cambridge, pp. 145-176.

Brill, Eric. "A simple rule-based part of speech tagger", *Proceedings of the Third Conference on Applied Natural Language Processing* (Trento, Italy), 1992, pp. 152-155.

Nicoletta & Remo Bindi. "Acquisition of Lexical Information from a Large Textual Italian Corpus", in Hans Karlgren (ed.), *Proceedings of COLING'90* (Helsinki, August 1990), Vol 1, pp. 54-59.

Carlson, Lauri. "Aspect and Quantification". In Philip Tedeschi & Annie Zaenen (eds.), *Syntax and Semantics, Volume 14: Tense and Aspect*, Academic Press, 1981, pp. 31-64.

Catford, J.C. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*, Oxford University Press, 1967.

Cherry, Lorinda L. "PARTS - A System for Assigning Word Classes to English Text", Computer Science Technical Report #81, Bell Lab., Murray Hill, N.J., 1978.

Chesterman, Andrew. *Contrastive functional analysis*. Amsterdam: Benjamins, 1998.

Church, Kenneth Ward. "A stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", *Proceedings of the Second Conference on Applied Natural Language Processing* (ACL), 1988, pp. 136-143.

Church, Kenneth & William Gale. "Inverse document frequency (IDF): a measure of deviations from Poisson". In David Yarowsky & Kenneth Church (eds.), *Proceedings of the Third Workshop on Very Large Corpora* (Cambridge, MA, EUA, 30 June 1995), 1995a, pp. 121-130.

Church, Kenneth & William Gale. "Poisson mixtures", *Journal of Natural Language Engineering* **1**, 2, 1995b, pp. 163-190.

Church, Kenneth & Patrick Hanks. "Word Association Norms, Mutual Information and Lexicography", *Computational Linguistics* **16**, 1, 1991, pp. 22-29, 1991.

Cruse, Alan. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford. Oxford University Press, 2004.

DeRose, Stephen J. "Grammatical category disambiguation by statistical optimization", *Computational Linguistics* **14**, 1, Jan. 1988, pp. 31-39.

Dixon, R.M.W. “A method of semantic description”, in Danny D. Steinberg & Leon A. Jakobovits (eds), *Semantics: An interdisciplinary reader in philosophy, linguistics and philosophy*, Cambridge: Cambridge University Press, 1971, pp. 436-471.

Dorow, Beate. “A Graph Model for Words and their Meanings”. PhD Thesis, IMS, Stuttgart University, 2006. http://elib.uni-stuttgart.de/opus/volltexte/2007/2985/pdf/diss_27022007.pdf.

Dunning, Ted. “Accurate Methods for the Statistics of Surprise and Coincidence”, *Computational Linguistics* **19**, Number 1, March 1993, pp. 61-74.

Edmonds, Philip & Graeme Hirst. “Reconciling fine-grained lexical knowledge and coarse-grained ontologies in the representation of near-synonyms”. In *Proceedings of the Workshop on Semantic Approximation, Granularity, and Vagueness (KR-2000)*, Breckenridge, Colorado, 2000.

Ellegård, Alvar. *The syntactic structure of English texts: a computer-based study of four kinds of text in the Brown University corpus*. Gothenburg: Acta Universitatis Gothoburgensis, 1970.

Ellis, John M. *Language, Thought and Logi*c. Evanston, IL: Northwestern University Press, 1993.

Fellbaum, Christiane (ed.). *WordNet: An Electronic Lexical Database*, with a preface by George Miller. The MIT Press, May 1998.

Garside, Roger, Geoffrey Leech & Geoffrey Sampson. *The Computational Analysis of English: A Corpus-Based Approach*, Longman, 1987.

Goodman, Nelson. "Seven strictures on similarity”. In Nelson Goodman (ed.), Problems and projects. Indianapolis, IN: Bobbs-Merrill, 1972, pp. 437-447.

Gonçalo Oliveira, Hugo, Diana Santos & Paulo Gomes. "Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação”, *Linguamática* **2**, 1, April 2010, pp. 77-94.

Green, T. R. G. “The Necessity of Syntax Markers: Two Experiments with Artificial Languages”, *Journal of Verbal Learning and Verbal Behavior* **18**, 4, Aug 1979, pp. 481-496.

Greene, Barbara B. & Gerald M. Rubin. “Automated Grammatical Tagging of English”. Providence, R.I.: Department of Linguistics, Brown University, 1971.

Grefenstette, Gregory. *Explorations in Automatic Thesaurus Discovery*. Boston, MA: Kluwer Academic Press, 1994.

Grefenstette, Gregory & Pasi Tapanainen. “What is a word, What is a sentence? Problems of Tokenization”, *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX'94)*, 1994, pp. 79-87.

Gruber, Thomas R. “A Translation Approach to Portable Ontology Specifications”, *Knowledge Acquisition* **5**(2), 1993, pp. 199-220.

Halliday, M.A.K. *Computational and Quantitative Studies*, vol 7 in the Collected Works of MAK Halliday edited by Jonathan J. Webster, London, New York: Continuum, 2005.

He, Ying & Mehmet Kayaalp. “A Comparison of 13 Tokenizers on MEDLINE”, Technical Report LHNCBC-TR-2006-003. http://lhncbc.nlm.nih.gov/lhc/docs/reports/2006/tr2006003.pdf

Heiden, Serge. “Interface hypertextuelle à un espace de cooccurrences: implémentation dans Weblex”, in Gérard Purnelle, Cédrick Fairon & Anne Dister (eds.), *7ième Journées internationales d'Analyse Statistique des Données Textuelles (JADT'04) "Le poids des mots" 10 - 12 Mars 2004*, vol 1, Presses Universitaires de Louvain, Louvain-la-Neuve, Belgique, 2004, pp. 577-588.

Hindle, Donald. “Acquiring Disambiguation Rules from Text”, *Proceedings of ACL 1989*, pp. 118-125.

Hindle, Donald & Mats Rooth. “Structural Ambiguity and Lexical Relations”, *Computational Linguistics* **19**, 1, March 1993, pp. 103-120.

Hirst, Graeme. “Ontology and the lexicon”. In Steffen Staab & Rudi Studer (eds.). *Handbook on ontologies*, Springer, 2004, pp. 209-229.

Jurafsky, Daniel & James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.

Justeson, John S. & Slava M. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering* **1**, 1995, pp. 9-27.

Katz, Slava M. "Distribution of content words and phrases in text and language modelling", *Natural Language Engineering* **2,** 1996, pp.15-59.

Kennedy, Graeme. "Over *once* lightly". In Carol E. Percy, Charles F. Meyer & Ian Lancashire (eds.), *Synchronic corpus linguistics: papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*, Rodopi, Amsterdam – Atlanta, GA, 1996, pp. 253-62.

Kilgarriff, Adam. "Which words are particularly characteristic of a text? A survey of statistical approaches", *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition* (Sussex, April 1996), pp. 33-40.

Kilgarriff, Adam. ""I don't believe in word senses"", *Computers and the Humanities* **31** (2), 1997, pp. 91-113.

Kilgarriff, Adam. "Language is never ever ever random", *Corpus Linguistics and Linguistic Theory* **1**, 2, 2005, pp. 263-276.

Kilkki, Kalevi. "A practical model for analyzing long tails", *First Monday* **12**, 5, May 2007, http://www.firstmonday.org/issues/issue12_5/kilkki/

Klein, Sheldon & Robert F. Simmons. "A computational approach to grammatical coding of English words", *Journal of the Association for Computing Machinery* **10**: 334-347.

Krenn, Brigitte & Christer Samuelsson. "The Linguist's Guide to Statistics: DON'T PANIC", 21 May 1997.

Macklovitch, Elliott. "Where the Tagger Falters", *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation* (Montréal, June 25-27, 1992), pp. 113-126.

Manning, Chris & Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, May 1999.

Marshall, I. "Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB Corpus", *Computers in the Humanities* **17** (1983), pp. 139-150.

Martins, João Pavão. *Knowledge Representation*. IST. http://www.cse.buffalo.edu/~rapaport/663/S02/martinssneps.pdf

Medeiros, José Carlos. "Avaliação de Correctores Ortográficos", *Actas do XI Encontro da Associação Portuguesa de Linguística*, Lisbon: Colibri, 1996, pp. 73-91.

Medeiros, José Carlos, Rui Marques & Diana Santos. "Português Quantitativo", *Actas do 1.o Encontro de Processamento de Língua Portuguesa (Escrita e Falada) - EPLP'93* (Lisbon, 25-26 February 1993), 1993, pp. 33-38.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller. "Introduction to WordNet: An On-line Lexical Database" (revised August 1993), *Five papers on WordNet*, 1993, pp. 1-25.

Monachini, Monica & Nicoletta Calzolari. "Standardization in the lexicon". In Hans van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht/Boston/London: Kluwer Academic Publishers, 1999, pp. 149-174.

Mosteller, Frederick & David L. Wallace. *Inference and Disputed Authorship*. 1964.

Nicolaeva, T. M. “Soviet Developments in Machine Translation: Russian Sentence Analysis”, *Mechanical Translation* **5**, 2, November 1958, pp. 51-59.

Pinker, Steven. *The stuff of thought: Language as a Window into Human Nature.* Allen Lane, 2007.

Pym, Anthony. *Translation and Text Transfer. An Essay on the Principles of Intercultural Communication*. Frankfurt am Main, Berlin, Bern, New York, Paris, Vienna: Peter Lang, 1992. Revised online version, Tarragona: Intercultural Studies Group, 2010, http://www.tinet.cat/~apym/publications/TTT_2010.pdf

Richardson, Stephen. “Determining Similarity and Inferring Relations in a Lexical Knowledge Base”, Ph.D. thesis, The City University of New York, 1997, Microsoft Research Report MSR-TR-97-02, ftp://ftp.research.microsoft.com/pub/tr/tr-97-02.doc.

Rivenc, Paul. “Vocabulário frequente e vocabulário disponível”. In Bacelar do Nascimento, Maria Fernanda, Paul Rivenc & Maria Luísa Segura da Cruz. *Português Fundamental*, Volume II, *Métodos e Documentos*, tomo 2, *Inquérito de Disponibilidade*, Lisbon: Centro de Linguística de Universidade de Lisboa, 1987, pp. 3-26.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson & Jan Scheffczyk. *FrameNet II: Extended Theory and Practice*. Printed June 15, 2010. http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126

Sampson, Geoffrey. *English for the Computer: The SUSANNE Corpus and Analytic Scheme.* Clarendon Press, Oxford, 1995.

Sampson, G.R. “Review of Christiane Fellbaum (ed.), *Wordnet: An Electronic Lexical Database*, 1998”, *International Journal of Lexicography* **13**, 2000, pp. 54-59.

Sampson, Geoffrey. *The ‘Language Instinct’ Debate*. March 2005. London & New York: Continuum International. Enlarged and revised edition of *Educating Eve*, Cassell, 1997.

Santos, Diana. *Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems*. Amsterdam/New York, NY: Rodopi, 2004.

Santos, Diana. “What is natural language? Differences compared to artificial languages, and consequences for natural language processing”. Invited lecture, SBLP2006 and PROPOR'2006, Itatiaia, RJ, Brazil, 15 May 2006. http://www.linguateca.pt/Diana/download/SantosPalestraSBLPPropor2006.pdf

Santos, Diana & Caroline Gasperin. “Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation”. In Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, 2002, pp. 597-604.

Santos, Diana, Luís Costa & Paulo Rocha. “Cooperatively evaluating Portuguese morphology”. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings*, Springer Verlag, 2003, pp. 259-266.

Saussure, Ferdinand de. *Cours de Linguistique Générale*, publié par Charles Bally & Albert Sechehaye avec la collaboration de Albert Riedlinger. Payot, Paris, 1972. First edition: 1915.

Sinclair, John. “Corpus Evidence in Language Description”. In Wichmann, Anne, Steven Fligelstone, Tony McEnery & Gerry Knowles (eds.), *Teaching and language corpora*. London & New York, Longman, 1997, pp. 27-39.

Snell-Hornby, Mary. *Verb-descriptivity in German and English: A contrastive study in semantic fields*, Carl Winter Universitätsverlag, Heidelberg, 1983.

Sovran, Tamar. "Between similarity and sameness”, *Journal of Pragmatics* **18**, 4, 1992, pp. 329-344.

Sparck Jones, Karen. “What's new about the Semantic Web?: some questions”, *ACM SIGIR Forum* **38**, 2, December 2004, COLUMN: Invited talks, pp. 18-23.

Steiner, George. *After Babel: aspects of language and translation*. Oxford: Oxford University Press, 1992 (1st edition 1975).

Stolz, Walter S., Percy H. Tannenbaum & Frederick V. Carstensen. "A stochastic approach to the grammatical coding of English", *Communications of the ACM* **8**, 6, June 1965, pp. 399-405.

Talmy, Leonard. "How language structures space". In H. Pick & L. Acredolo (eds.), *Spatial orientation: theory, research, and application*. New York, Plenum Press, 1983, pp. 225-282.

Tversky, Amos. "Features of similarity", *Psychological Review* **84**, pp. 327-352.

Underwood, Nancy, Patrizia Paggio & Gurli Rohde. "A methodology for evaluating Spelling Checker functionality: Developing test suites for Danish", in Kimmo Koskenniemi (ed.), *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics* (Helsinki, 29-30th May 1995), 1995, pp. 76-85.

Veale, Tony. "Enriched Lexical Ontologies: Adding new knowledge and new scope to old linguistic resources", ESSLLI 2007, Dublin, http://afflatus.ucd.ie/papers/Essilli_EnrichedLexiOnto.pdf

Wilks, Yorick. "Is Word Sense Disambiguation Just One More NLP Task?", *Computers and the Humanities* **34**, 1-2, April 2000, pp. 235-243.

Wilks, Yorick & John Tait. "A Retrospective view of Synonymy and Semantic Classification". In John I. Tait (ed.), *Charting a New Course: Natural Language Processing and Information Retrieval: Essays in Honour of Karen Spärck Jones*, Springer, 2005, pp. 255-282.

Wilks, Yorick & Christopher Brewster. "Natural Language Processing as a Foundation of the Semantic Web", *Foundations and Trends in Web Science* **1**, 3, March 2009, pp. 199-327.

**Image credits**

Contar carneiros: http://www.oasrs.org/conteudo/agenda/noticias-detalhe.asp?noticia=1549 (obtained 4 June 2010).

Topic maps: http://en.wikipedia.org/wiki/File:TopicMapKeyConcepts2.PNG (obtained 16 June 2010)

Expectations: Joakim Krøvel, Scanpix, www.scanpix.no

**Personal acknowledgements**

Diana Santos thanks Danilo Giampiccolo for help in tracing Bortolini et al. (1981) right from Italy, Eric Atwell for permission to use an interesting email exchange, back in February 2010, Anton Landmark for the Humpty Dumpty quotation and for encouragement and criticism, Nuno Cardoso, Tormod Håvaldsrud and all other colleagues at SINTEF who attended a preliminary version of this course, Doris Lund from the SINTEF library for granting access to a vast number of books and papers, João Pavão Martins for the SNePs figure, Maarten Marx for permission to use his site and photographs therein, and last but not least, acknowledges financial support in the scope of the Linguateca project, co-financed by the Portuguese government, the European Union (FEDER and FSE) under POSC/339/1.3/C/NAC contract, UMIC and FCCN.

## Words and their secrets: Introduction

Diana Santos & Maria José Bocorny Finatto
ESSLLI 2010



## Course preview

- Monday
  - Introduction (D+MJ)
  - Linguistic evolution: from words in the mind to real utterances (MJ)
- Tuesday:
  - Basic technologies: spell checking and POS tagging (D)
  - Word types and their function in texts (MJ)
- Wednesday
  - Dictionaries, lexical networks, lexical ontologies, wordnets and wordclouds (D)
  - Lexicography and terminography: old traditions and new routes (MJ)
- Thursday
  - Frequency studies in Portuguese: *de* and *Brasil* (MJ)
  - Lexical statistics (D)
- Friday
  - Vagueness, ambiguity, and multilingual issues (D)
  - Conclusions (MJ+D)



## One word (*Duas palavrinhas...*) about the lecturers

- **Diana Santos** has worked in natural language processing of Portuguese for 25 years. She is the leader of the *Linguateca* project, an international network for resources and evaluation for the Portuguese language (1998-). Her PhD in 1996 was on corpus-based semantic studies. She is a researcher at SINTEF (Norway) and FCCN (Portugal), in Oslo
- **Maria José Bocorny Finatto** has worked in Terminology, Lexicography and Linguistics of Brazilian Portuguese for 20 years. Her PhD in 2001 was on terminology and the specific theme was definitions patterns in Chemistry dictionaries and texts. She is a researcher at Federal University of Rio Grande do Sul, located in Porto Alegre, a South Brazil city
- They met in EBRALC, Nov 2008 in São José do Rio Preto, Brazil



## Saussure and two different things

- *Historiquement la négation **pas** est identique au substantif **pas**, tandis que, pris dans la langue d'aujourd'hui, ces deux éléments sont parfaitement distincts.* (Saussure, 1916: 129)

  'Historically *pas* as negation is the same as the noun *pas* ['step', DMS], while in today's language, these two elements are perfectly distinct'
- Saussure was the first to emphasize that the researcher/analyst has different facts/different words if s/he is studying synchronic linguistics or diachronic linguistics



## Word senses

- Traditionally and as reflected in dictionaries, a word may have more than one sense: *bank*, *hand* and *cerca* are well-known examples... But, **what is a word sense**?

  *Among other possibilities, a word sense may be regarded as a purely **mental object**; or as a structure of some kind of **primitive units of meaning**; or as the **set of all the things in the world that it may denote**; or as a **prototype** that other objects resemble to a greater or lesser degree; or as an **intension** or **description** or **identification procedure** [...] of all the things that the sense may denote* (Hirst, 2004: 214)



## Words are not only the realm of linguistics, literature, or journalism...

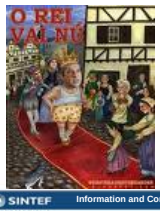
- Take anthropology:

  *[...] paradoxical quest: how to translate untranslatable phrases and words. [...] **words are the elements of speech, but words do not exist.** Having once recognised that words have no independent existence in the actual reality of speech, [...] the intermediate link between word and context, the linguistic text.* (Malinowski, 1935:23)

  Malinowski, Bronislaw. *Coral gardens and their Magic: A Study of the Methods of Tilling the Soil and of Agricultural Rites in the Triobrand Islands. Vol 2. The Language of Magic and Gardening.* New York: American Book Company, 1935.

## Pictures make no sense without words

- Why has this picture been taken? Where? To illustrate what? To be used as a joke, a report, or a work of art? What events or feelings is it meant to invoke in the seer?



How my students end up....

http://staff.science.uva.nl/~marx/

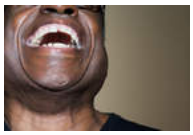


## The cultural dependence of captions

- *Man reading*. This can be a good enough caption in an European museological context, but certainly not in an Asian, or African context



## Middle aged-man ... man laughing



- interaction with unrelated subjects or issues



## "Same" concept: Foxes and blue

Figure of fox and the prince, from *Le petit prince* by Antoine de Saint Exupery

Figure of fox from *Klatremus og de andre dyrene i Hakkebakkeskogen* by Thorbjørn Egner

- Feminine, related to love and friendship
- Perfect: *Ouro sobre* ***azul***

- Masculine, related to tricks
- Depression: the blues



## *O sexo dos anjos* (lit: the gender of angels)

- This (pointless) is best translatable by "splitting hairs" in English, because English has no gender as a grammatical category
- But for languages which do: *I actually chose this example because it sparked an unexpected flurry of debate on the corpus.quran.com feedback blog; it turned out that Arabic "angel" nearly always has male gender, but there are a couple of cases where the gender affix is female, but some quranic scholars maintain that even these cases are* ***gender*** *rather than* ***physical*** *sex, and that god and all angels must be male – they cannot accept that some leading higher beings could be female. I've avoided getting involved in this debate but it was interesting seeing the amount of time and effort put into the blog...* (Eric Atwell, 9 Feb 2010, personal comm.)



## Abstract concepts

- Expectations

**Joakim Krøvel at Scanpix**

## Words are the units of categorization

- And for this they are equally important to philosophy and to logic
- Words are the natural units of an extremely complex classification system which is a natural language. There are subunits (morphology, letters, sounds, ...) and superunits (multiword expressions, phrases, sentences, texts, turns...) but words are the most basic. In other words, they are the organizing thread, the popular knowledge, the more difficult to define (the other categories can be described resorting to the basic notion of word)
- In a cline with grammatical categories and features and grammar proper, words are specific to a specific language system



## What is the relationship between words and reality?

- What is the relationship between language and reality?
- The naive view(s):
  - Direct reference
  - Ostensive denotation
  - Harwired in the brain (mental images)
- More enlightened views/aproaches
  - Words (and the other mechanisms of language) represent classes of different objects which are considered, for the purpose of conceptualization, as similar
  - Words are a prerequisite for thought and communication
  - Words (and the rest of language, including communication patterns) are learned through interaction with the language community (and especially the mother)



## A common conceptualization

The triangle: mind, language, context/reality

- There is no language without mind, and the mind is always in a context/reality, so we should perhaps say **reality as perceived by the mind/senses/soul** (and this is an internal, private, personal matter) and language as a social system and as perceived by the person (again, each internal language is one's idiolect)
- As again Saussure pinpointed, language is **arbitrary** in the sense that there is NO reason for the particular signs, but it is **obligatory** because no one can change it individually, since it pressuposes a social contract: premises for the successful use of a language is that (a) one conforms to the rules and (b) one teaches the (arbitrary) rules to one's children/co-citizens



## What does it mean to <u>know</u> a word?

- To be able to pronounce/spell it correctly?
- To be able to use it felicitously?
- To be able to define it?
- To be able to provide synonyms or near synonyms?
- To know its history / etimology?
- To know its morphology?
- To know its social consequences?
- To be able to provide translation into another language?
- To be able to point to the instance denoted by the word?



## What is a word? First answer(s)

- Meaningful building blocks of a language
- Corresponding to what people thought worthy of naming
- Constantly fluctuating and acquiring new meanings and losing old meanings
- Always working in context
- Always related to situation and co-text
  - Of course one can devise non-words, that is, sequences of sounds or letters for psycholinguistic or medical tests
- But
  - One can reify a word and provide a definition for it
  - One can use the word as an abstract term for all possible denotandums



## How does a language choose its units?

Talmy's (1983:277ff) suggests:

- *The majority of semantic domains in language are* n*-dimensional, with* n *a very large number. For example, no fewer that [] twenty parameters are relevant to the domain of spatial configuration as expressed by closed-class elements such as English prepositions and deictics. [List]*
- *With so many parameters, full domain coverage by fairly specific references would require thousands of distinct vocabulary items, […]*
- *Rather that a contiguous array of specific references, languages instead exhibit a* ***smaller number*** *of such references in a* ***scattered distribution*** *over a semantic domain. That is, a fairly specific reference generally does not have any immediate neighbors of equal specificity.*

## How are the different levels of language related?

- Difference between closed class/grammatical words, and open class words
- Difference between inflection, derivation, and other satellites
- Difference between sentence, clause, phrase and word
- Difference between what is implicit, default, unsaid, taboo...

- One of the purposes of this course is to present many of the different answers that have been given to these questions!

SINTEF Information and Communication Technologies

Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil UFRGS 19

## Basic technologies: Spellchecking and POS tagging

Diana Santos

**WATS: Words and their secrets**, ESSLLI 2010
Diana Santos & Maria José Bocorny Finatto

Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil

WATS:: Basic technologies: spell checking and POS tagging 1

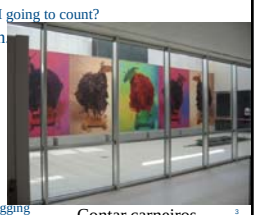
## Preview

- Word count and the type/token distinction
- A simple word-based technology: spelling verification and correction
  - Overview and history
  - Challenges
- Another simple word-based technology: part-of-speech tagging
  - Overview and history
  - Challenges

WATS:: Basic technologies: spell checking and POS tagging 2

## Types and tokens

- Presuppositions of counting: Individuation, and classification
- In order to count, one has to abstract from differences and assign the same label to different individuals
- How many people are in this room?
  - First, which of the objects in the room am I going to count?
- How many native languages are spoken...
  - Assign native language to each
  - Different native languages: different types
- How many languages are spoken ...?
  - Assign languages to each person
  - Different languages: different types



Contar carneiros

WATS:: Basic technologies: spell checking and POS tagging 3

## Type/token ratio

- The type/token ratio is therefore something that depends on the classification the researcher/counter is interested in
- When one talks about word type/token ratio...
  - One may be classifying words just by form
  - One may be classifying words by the lexical paradigm they belong to
    - Lemma
    - Capitalization, ortography
    - Lemma and POS
    - Meaning
    - ...
- Exercise: *This is a dubious example of this sort of thing, that provides different values if differently computed with different sorts of computing using example things of dubious computing value.*

WATS:: Basic technologies: spell checking and POS tagging 4

## Spell checking

- Identify / detect incorrectly spelt words
- Suggest corrections
- Automatically correct
- Words are defined as sequences of "word-proper" characters, separated by word separators
- "Incorrectly" spelt means
  - not belong in the dictionary
  - not being accepted by a set of (language-specific) given rules
  - not numbers or simple letters

WATS:: Basic technologies: spell checking and POS tagging 5

## Issues in spell checking

- What to encode in the dictionary?
- Rare words may correspond to errors
- Some of the most frequent errors (exchange between false friends) can only be detected in context:
  - its / it's (en)
  - å / og (no)
  - à / a (pt)
  - two / to (en)
- What if the error is absence or addition of word-separator?
  - callback/ call back (this problem is compounded in languages with compounds)
  - Fee dback
- Avoid correction of (some) proper names

WATS:: Basic technologies: spell checking and POS tagging 6

## Spelling correction

- How to evaluate/rank the best suggestions?
- How to provide measures to compare different spellcheckers?
- Number of correct corrections/Number of (first) corrections suggested
- Number of correct corrections/Number of errors

*I don't likke cracrashes*

- How to count the number of errors? Words with errors?
- And how to count the number of correct suggestions if the number of words can be different after correction?

- There are incorrect corrections which are nevertheless useful!



## Further examples

- Dirigi lhe — dirigi-lhe — 2 or 1 words / 2 or 1 errors
- Senti-la-hia — senti-la-ia — 3 or 1 words / 1 error
- Ta, to
- 'Tás, 'tamos — estás, estamos — errors?
- diversidade.Nesse — diversidade. Nesse — 1,2 or 3 words?
- dêmo — dê-mo — 2 or 1 words / 1 error
- auto-denominado-se — auto-denominando-se
- PhG — PhD
- rock'n'roll, Toys'R'Us, 90's, M'Gladbach, 2000-2010, ...
- R&D, A4, UB40,



## Further examples (2)

- Momentos-«Chávez»
- Ex-comandante da LUAR
- Pré-25 de Abril
- Pós-11 de Setembro
- Decreto 3.048/99, (0xx21)2550-9268, (0xx21) 2550-9268
- Av. Tucunaré, 720 - Tamboré - CEP 06460-020 - Barueri – SP
- e/ou, and/or
- q.b.
- telemóveis 3G
- PNALE 2005-2007



## Quantitative summing up

- Take a small corpus created for NER evaluation in Portuguese, with 129 texts, in the scope of HAREM
- Input to it as txt, Word considers **78,832** words
- In Linux, `wc -w` states **78,825**
- After parsing with PALAVRAS a broad-coverage parser for Portuguese (Bick, 2000), we got **88,911 tokens**, **84,455** words
- After applying AC/DC tokenization fixes, we end up with **85,978** tokens, of which **80,391** are considered words
- Differences in tokens from the Linguateca tokenizer and PALAVRAS tokenizer: **16,055** differences
  - Only in the word forms:



## *Morfolimpíadas*: the tokenization nightmare

- In Santos et al. (2003) we reported on the preliminary results (trial run) of the *Morfolimpíadas* evaluation contest:
  *Even if all systems returned exactly the same analyses for the forms they agreed upon, there would still be disagreement for 15.9% of the tokens or 9.5% of the types*
- Four different systems
  - Common types, case 1: 8480
  - Common types, case 2: 9580

Case 1: Running text

| | | | | |
|---|---|---|---|---|
| No. of tokens | 41,636 | 41,433 | 39,503 | 41,197 |
| Common tokens | 84.1% | 91.6% | 86.5% | 86.2% |

Case 2: List of different tokens

| | | | | |
|---|---|---|---|---|
| No. of types | 11,593 | 10,896 | 10,613 | 10,745 |
| Common types | 90.7% | 92.0% | 91.3% | 90.5% |



## Thirteen tokenizers (He & Kayaalp, 2006)

- 18, down to 13, freely available software packages
- for use in MEDLINE abstracts
- Test with 78 MEDLINE abstracts
- Number of tokens varies from 14,488 to 17,117
- Rough evaluation criteria
  - Source code available, and programming language it is writtten in
  - Merging/losing original information (bracket kind, words etc.)
  - Compound words
  - Words mixing letters and numbers
  - Inconsistencies
  - Codepage, such as Unicode, supported

## Evaluation of spellcheckers (Medeiros, 1996)

Medeiros (1996) suggested three kinds of evaluation axes

- Processing speed (verification speed; average answer time; suggestions / second)
- Functionality
- Result **accuracy**
  - suggestion dispersion (sugg./word; sugg. factor)
  - suggestion ordering (ordering factor)
  - failure (missing, zero, completeness, robustness)

$F_i=N_f/N_s$ $F_z = N_z/N_e$ $F_c=1-N_d/N$ $F_r=1-N_{in}/N_e$



## Evaluation of spellcheckers (cont.)

- Correction index $\frac{(F_i + F_z)}{(F_d + F_c + 2F_r)}$
- Test materials
  - Free text
  - List of errors
  - List of pairs (error, correction)
  - List of error-free words ordered by frequency in real-text corpora (Underwood et al., 1995)
- Exercise: spellcheck a small text in the language you are interested in, produce different values for the number of words/errors, and see how this can affect the different measures for the correction index



## PoS tagging

- Apparently the easiest and best defined task...
  - Manual or manually validated vs. automatic
- For each **word**, assign the correct part-of-speech
- Word?
  - And multiwords? And named entities? And non-words?
- For each word, assign the **correct** part-of-speech
  - Correct? Depends on the theory of grammar
  - Only one tag?
  - Evaluation of PoS tagging... what is correct?
- For each word, assign the correct **part-of-speech (PoS)**
  - Only PoS? Or morphology as well? Or subcategorization? Or everything?



## Some history of POS tagging

- Apparently the first machine disambiguation of natural language text was done by Russian researchers working on MT (Nicolaeva, 1958)
- Klein & Simmons (1962) develop a *first component in a syntactic analysis program, which is part of a larger QA system*
- Stolz et al. (1965) *apply statistical methods: decisions ... based on conditional probabilities of various form classes in given syntactic environments* -- Cherry (1978) *assigns part of speech by rule*
- Green & Rubin (1971) create the first annotated corpus, the Brown corpus, human revised; and Ellegård (1970) the first human annotation
- Marshall (1983) improving LOB based on Brown
- DeRose (1988), Church (1988), Garside et al. (1987), Hindle (1989): POS tagging as pre-processing



## Macklovitch (1992): First linguistic analysis?

- *Generally speaking, a given tag set may be more or less suitable for certain applications*
  - *after, before, until* can be either IN or CS
  - *analogous to suppressing the distinction between verbs that subcategorize for an NP or for a sentence...*
  - *-ed* forms can be either VBD or VBN
  - nouns or adjectives: JJ or NN
- Global dependencies (instead of "long-distance"): whether a verb is in imperative, present or subjunctive can depend on the whole sentence.
- *Why bother*?
  - Evaluation relevance
  - Automatic error detection – and maybe even correction



## Brill tagger (1992) learns from its weaknesses

- "A simple Rule-Based PoS Tagger": *robust and rules automatically acquired*
- Currently called a hybrid method, because it uses **machine** learning, but requiring **human** annotated data
- First it assigns the most frequent tag to the already existing words in the training material; then uses the word endings out of the dictionary
- Comparing the output to human annotated material, it creates error triples: <old category, new category, frequency>
- Eight different patches are tried out, and the one which provides higher global error diminishing is added to the patch list
- 71 patches, 5% error in 5% of the Brown corpus

## Measuring Portuguese POS ambiguity

- Medeiros et al. (1993): potential word classes in a corpus
  - **n/a**, **vpp**, **v**, adv, pf, cl
  - 1.02494 classifications per form; 1.1398 class/form if only the three first are considered
- Bacelar do Nascimento et al. (1993): real word classes in a corpus
  From a corpus of transcribed oral speech, 700,000 words (25,107 types), reduced to the forms corresponding to lemmas with frequency > 40 (1553 lemmas): 65,000 forms, where there were potentially 834 ambiguous lemmas, corresponding to 1371 POS-ambiguous form (types), whose occurrences were then analysed in context
  - N-ADJ: 143 types: 123 Noun, 121 Adj
  - N-ADJ-V: 66 types: 44 Noun, 57 Adj., 35 Verb



## Kennedy about the value of POS tagging

- *When claims are made about the impressive accuracy with which grammatical tags can be assigned by machine, it is often not made clear to consumers that the* ***high success rates [] are based on an averaging process****. [] certain very frequent words or word classes can be tagged with virtually total accuracy, while for other items, accuracy rates of 80-85% are more typical.* (Kennedy, 1996:253)
- 100 most frequent word types in LOB -> 49% of the tokens
- Ca. 2/3 (65, types, ca. 335,000 tokens) belong to one class only!



## Is POS to be evaluated presupposing correct lemma attribution? Or do the two tasks go hand in hand?

- Of course in some cases they do: different lemma, different POS
  - *desses* (dar, V, or desse, PRON) "you would give", "from those"
  - *suas* (suar, V, or seu, PRON) "you perspire", "his or hers"
  - *era* (ser, V, era, N) "was", "era"
- But in others they don't: same lemma, different POS
  - *creme* (creme, ADJ, creme N) "beige", "cream"
  - *alto* (alto, N, alto, ADJ, alto, ADV) "tall", "loudly", "top"
- But in others they don't: same POS, different lemma
  - *costas* (costa, N, costas, N) "coasts", "back"
  - *assente* (assentar, V, assentir, V, ... assente, ADJ) "write down", "agree"...
  - *fora* (ser, V, ir, V, ... fora, ADV) "had been", "had gone", "outside"
  - *vendo* (ver, V, vender, V, vendar, V) "see", "sell", "cover (eyes)"



## What is the right POS? (Santos & Gasperin 2002)

- This question is always **according to a particular theory of grammar**
- What's the use of providing the same POS for different syntactic constructs?
  - *Ele está de volta* (he is back)
  - *De volta da mãe, ele apressava-se* (around his mother, he hurried)
  - *Comprou o bilhete de volta* (s/he bought the return ticket)
- They could be obviously separated by
  - Tokenization ("de volta", "de volta de", "volta", "bilhete de volta")
  - Syntactic function (predicative, adjunct, specifier?)
  - Syntactic constituent they head/belong (PP, AVP, NP)



## Concluding remarks

- Beware of "easy" tasks, light hearted procedures
- Even for the least intelectually challenging task... Criteria for "wordness" have to be thought and decided upon.
  *In linguistic textbooks tokenization is quickly dispatched as a relatively uninteresting pre-processing step performed before linguistic analysis is undertaken. In reality, tokenization is a non-trivial problem* (Grefenstette & Tapanainen, 1994)
- In the next days this will be shown in other fields on natural language processing as well ...

## Dictionaries, lexical networks, lexical ontologies, wordnets and wordclouds

Diana Santos

**WATS: Words and their secrets**, ESSLLI 2010
Diana Santos & Maria José Bocorny Finatto

Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil



## Preview

- Dictionaries... And frequency dictionaries, and fundamental vocabularies
- Terminology and history: network, ontology, wordnet, word cloud...
- What are nodes? What are lexical relations? What is the purpose of linking nodes?
- How are words defined by the (network) company they keep?
- Examples from several "schools"
  - Inheritance in LKBs
  - Dorow: topology
  - Classical AI semantic networks
  - Hirst: Near synonyms
  - Miller: Synsets
  - FrameNet



## First: how to choose items for a frequency dictionary

- Frequent words? Corpus-based: how to choose the corpus?
- Frequency of lemmas (implies lemmatization and corpus analysis), or of forms?
- Frequency is not the only thing that matters: Dispersion, repartition, frequency stability...
- Provided the corpus is subdivided in **n** parts (it is possible to subdivide it), and one has $f_1$, $f_2$, ... $f_n$ and thus $f_1$-$f_{av}$, $f_2$-$f_{av}$, etc.
- Bortolini et al. (1981:21-30) suggest the formula proposed by Juilland & Chang Rodriguez for the *Frequency Dictionary of Spanish Words*, and use (n=5): first the 5,000 lemmas with higher U, then add all that have R>= 3 with the same U, then U=1.78, resulting in 5,352 lemmas

$U=FD$ $\quad D=1-S/f_{av}\sqrt{n-1}$ $\quad S^2=\sum(f_i-f_{av})^2$



## Fundamental vocabularies

- Frequency is not enough: what about availability?
- If *knife* is frequent, would not *fork* qualify as available – and therefore required to be included as well?
- If one knows how to use *bachelor*, one knows the meaning of *married*
- Rivenc (1987)

| | corpus | voc. | themes |
|---|---|---|---|
| Français fondamental | 312135 | 806 | 15 |
| Português fundamental | 700000 | 1179 | 27+3 |
| Español fundamental | 800000 | 949 | 25 |

- A list of themes/interest centers: elicitating words after a theme (human body, games, village, school, politics, ...).
- Threshold frequency: $F_1$ frequency of the highest ranked word, N the number of words requested, D is dispersion, K is a adjusted parameter

$F_L=N*K*F_1/D$



## Cobuild: dictionary/grammar for the people

- *... the intuitions about language which [fluent speakers]can access are substantially at variance with their own language behaviour* (Sinclair, 1997:29)
- A set of precepts for language teaching:
  - Present real examples only — people can borrow books according to...
  - Know your intuition
  - Inspect co-texts
    *it is difficult, in the face of the evidence, to continue to rely on the idea of each word deliverings its little nugget of meaning* — physical appearance
  - Teach by meaning
    *if a word has two meanings one can predict with confidence two structures at least*
  - Highlight productivity



## Lexical knowledge bases

- Inheritance networks used for semantics (Kilgarriff, 1995)
- Incorporate regular polysemy in the dictionary/LKB
- *words have an indefinite number of potential senses*
- tree/wood alternation
- tree/fruit alternation
- transitive alternations
- ...

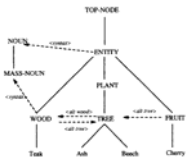
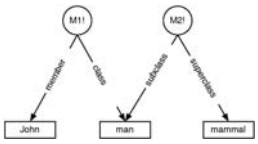


Figure 15.3. Taxonomy showing transitive alternations.

## Good old semantic networks

- Artificial intelligence knowledge representation: networks that allowed for extended reasoning around structured concepts
- SNePS, http://www.cse.buffalo.edu/sneps/
  - nodes represent intensional concepts
  - path-based reasoning




Martins (2002)



## The concept of network

- Gently...
- From just a convenient graphical representation for laymen ... to a mathematical discipline...
- Where is the use of networks located in language studies (including linguistics, natural language processing, and terminology...)?

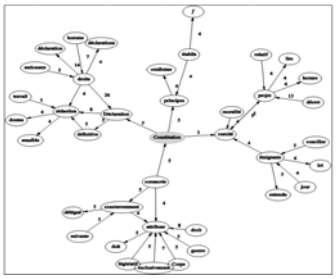
Excursus:

disciplines borrow freely from other disciplines so that, in the end, language is really natural language no matter special purpose languages are defined. A good example is *network*



## The word *Constitution* in speeches by Mirabeau (Heiden, 2004, fig. 13)

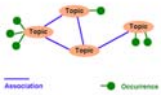
- Exploitation of a specific co-occurrence index in the scope of a hypertext computational environment, *Weblex*
- Exploration of a closed corpus (the speeches of the *Assemblée constituante*)
- Lexicometrics: recursive lexicogram





## Just a convenient graphical representation?

- Topic maps: *is a standard for the representation and interchange of knowledge, with an emphasis on the findability of information*
- *Topic Maps (vs RDF)*
  - *(i) provide a higher level of semantic abstraction (providing a template of topics, associations and occurrences, while RDF only provides a template of two arguments linked by one relationship)*
  - *and (hence) (ii) allow n-ary relationships between any number of nodes, while RDF is limited to triplets.*





Wikipedia (16/6/2010)



## Ontology vs. Lexicon (Hirst, 2004)

- Computational lexicons = vocabulary (=list of words) plus information on them
- Lexical entry = a large record, w/ inheritance and generative properties
- Word senses and semantic structure of the lexicon
- Lexicons are not (really) ontologies
  - Lexicons are linguistic objects
  - Ontologies aren't
- Lexically based ontologies and ontologically based lexicons
  - It depends on what the ontology is for
  - Covert categories: wear, things-that-carry-people, ...
  - If it is to deal with language(s)... machine translation, text understanding...



## What is an ontology?

- "Little o" versus "Capital o"
- Different definitions depending on the subject
- People divided by a common word (a pun on Shaw's)
- The main difference(s) seem(s) to be
  - Are instances in, or out?
  - Is there a difference between tokens and types?
  - Is there a difference between proper names and common names?
- What are concepts (and labels for them)?
  - Are they real?
  - Are they pseudo-labels? But really the words/terms naturally mean

#man

## What is an ontology, part 2

- Even if it is not explicit, it always includes **relations**
- Does it include reasoning rules?
- Does it also include elements that can be obtained by reasoning?

- In the informatics community, Gruber's (1993) definition is accepted: *An ontology is a formal explicit specification of a shared conceptualization for a domain of interest.*
- In the linguistic community, I propose Veale's (2007) definition of lexical ontology: *An ontology of lexical(-ized) concepts, used in NLP, serving as a lexical semantics* (ESSLLI 2007, Enriched Lexical Ontologies)



## What are domain ontologies?

`*When I use a word*,' Humpty Dumpty said, in rather a scornful tone, `*it means just what I choose it to mean -- neither more nor less.*'

- What are domains?
- Do the concepts in a domain mean separately?
- Do they only mean their place in the ontology?
- How are domains or genres defined?
- LSP, LSP, LSP or the need for power.
  - My term is better than yours
  - My school redefined all terms
  - My definition is better than yours
- Circularity or hermeneutics?
- What is **general** language?
- Is there an ontology for general language?



## Same or similar revisited

- Similarity is relative, variable, culture dependant (Goodman, 1972)
- Circumstances alter similarities (Goodman, 1972)
- The similarity of objects is modified by the manner in which they are classified (Tversky, 1977)
- "similarity" is a sign that is attributed to a set of entities, attributed by someone and also interpreted by someone (Chesterman, 1998)
  - similarity-as-trigger
  - similarity-as-attribution
- the greater the extension of the set of items assessed as being similar, the less the pertinent degree of similarity
- Tension between "oneness" and "separate individuation" (Sovran, 1992)



## Example from Tversky (1977)

**Question**: To which country is Austria more similar to?

- Sweden, Poland, Hungary

  *Sweden* (49%)
- Sweden, Norway, Hungary

  *Hungary* (60%)

Let us try again

- Germany, Denmark, The Netherlands
- Germany, Switzerland, The Netherlands



## Differences between languages

(1) *"I want a different apple." "Why? They are all the* ***same****."*
(2) *They wore the* ***same*** *dress.*
(3) *I'll have the* ***same*** *as her (said to a waiter).*
(4) *These two pens look* ***similar****, but one is more expensive than the other*

- English *same* is ambiguous between type and token identity

- Finnish: not the same item in (1) nor (2), but in (3).
- Portuguese: not the same item in (1): *são todas iguais*
- Portuguese: *parecem iguais* in (4)



## *A Graph Model for Words and their Meanings*

- PhD thesis by Beate Dorow, IMS, 2006
- Graph-theoretic approach to the automatic acquisition of word meanings
- [...] represent the nouns in a text in form of a semantic graph consisting of words (the nodes) and relationships between them (the links)
- Links in the graphs are based on cooccurrence of words in lists

## An ambiguous English word: *rock*

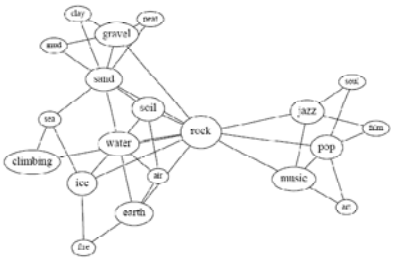
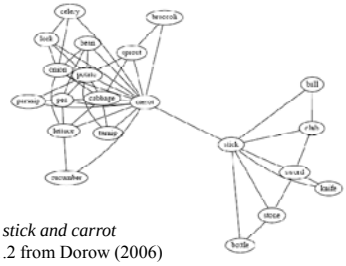



Fig. 1.1 from Dorow (2006)



## Idiomatically related words link different subgraphs




*stick and carrot*
Fig. 1.2 from Dorow (2006)



## Automatic detection of synonyms: sharing many neighbors


*author* and *writer*
Fig. 1.3 from Dorow (2006)



## Some details in Dorow (2006)

- Network of nouns built from POS-tagged English text (BNC)
- Which are connected by conjunctions *and*, *or*, and *nor*
- NP(, NP)*, (CJC NP)+
- NP defined as a sequence AT? CRD* ADJ* NOUN+
- Preprocessing: replace by WordNet base forms if unique
  - Cars -> car *cars* is replaced by *car*
  - Aids -> aid, aids *aids* is **not** replaced by *aid*
- Elimination of weak links: edges which do not occur in a triangle are eliminated
- Problems: POS errors, non-symmetry in lists, MWEs ☺

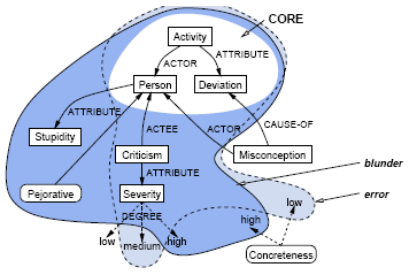


## The subtle problem of quasi-synonyms 1

- *Contrary to what one might expect – that the more similar two items are the easier it is to represent their differences (...) there is actually remarkable complexity in the differences bewteen near-synonyms* (Edmonds & Hirst, 2000)
- A model of fine-grained lexical knowledge
  - Core denotation, inherent context-independent, language-neutral
  - Peripheral concepts: structures of concepts defined in the same ontology as core denotations are defined in, used to represent non-necessary and indirect aspects of word meaning
  - Fr. (*faute*, *erreur*, *faux pas*, *bavure*, *impair*, *bêtise*, *bévue*), En. (*blunder*, *lapse*, *mistake*, *slip*, *howler*, *error*)
  - Fr. (*ordonner*, *commander*, *sommer*, *enjoindre*, *décréter*), En. (*command*, *order*, *bid*, *direct*, *enjoin*)



## Cluster of **error** nouns (Edmonds & Hirst, 2000)






SINTEF Natural Language Technologies Group
FCCN Fundação para a Computação Científica Nacional
UFRGS
Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil

## The $$ of plurals vs. singulars (Pinker 2007)

- Google sells index terms
  - In order for appropriate adverts to appear together with the results
  - "photo cameras" is more expensive than "photo camera"

  ... because it shows that people are undecided about which one to choose
- All conflation is of course reductive
  - *squashes* (En.) are ONLY vegetables, while *squash* is ambiguous (Dorow)
  - *pais* (Pt.) can mean parents as well as fathers (plural of *pai*)
  - Bindi et al. (1994) describe the need for observation of word forms
    - contatto (It.): *... Only three words out of twelve really apply to the lemma* ***contatto****. The other nine either co-occur with the singular or with the plural*



## WordNet

- WordNet started as psycholexicologist's model of word meaning
- (psycholexicology = research concerned with the lexical component of language)
- An On-line Lexical Databae
- The initial idea was to "provide an aid to use in searching dictionaries conceptually (...) to be used in close conjunction with an on-line dictionary of the conventional type"
- Miller et al. (1993): *a dictionary based on psycholinguistic principles*
  - expose (psycholinguistic) hypotheses to the full range of vocabulary
  - organize lexical information in terms of word meanings, rather than word forms



## WordNet ... and wordnets

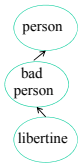
- One the most well-know and used lexical resources for English
- An example/model for several other languages
- A lot of wordnets and wordnet-alignment word, Global WordNet conferences all around the world

- Free for use, abundant computational support
- Several new developments/augmentations:
  - definitions, domains, addition of other sources, etc.
- But: are all uses warranted or appropriate? Is the underlying WordNet linguistic/semantic theory sound? Or applicable in every application?



## Sampson's (2000) critical remarks

- *it seems surprising that a database constructed manually by academics with no access to a dictionary-publisher's archive could be a seriouis contender as the leading tool in this domain*
- *… network of hyponymy relationships between nouns apparently requires some nodes which correspond to no single item of English*
- *The system is so naive that it (…) recognizes no distinction between the species/genus relationship, as in* horse/animal, *and the individual/universal relationship, as in* Shakespeare/author, *treating both indifferently as cases of "hyponymy"*





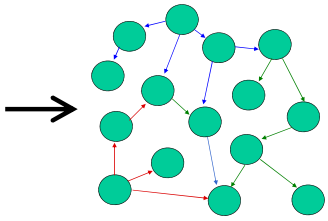

## WordNet and MindNet

- Automatic creation of a similar lexical network from the merge of (the parsing of) several dictionaries
- (Machine-readable) dictionary parsing
  - Calzolari for Italian
  - Amsler for British English, Chodorow for American English
  - Montemagni & Vandervende
  - Ide & Véronis
- (Machine-readable) dictionary using for parsing
  - Jensen & Binot
- MindNet: Microsoft Research lexical network
  - Fellbaum's discussion, Richardson's discussion



## PAPEL and its evaluation

- *Palavras Associadas Porto Editora – Linguateca* (Gonçalo Oliveira et al., 2010)
- http://www.linguateca.pt/PAPEL/

## Sowa's conceptual graphs

**Conceptual Graph (textual)**

- [Cat: ∀]->(On)->[Mat].
- [Go]-
  - (Agnt)->[Person: John]
  - (Dest)->[City: Boston]
  - (Inst)->[Bus].
- [Person: Tom]
  - <-(Expr)<-[Believe]->(Thme)-[Proposition: [Person: Mary *x]
  - <-(Expr)<-[Want]->(Thme)-[Situation: [?x]<-(Agnt)<-[Marry]->(Thme)->[Sailor] ]].

**Natural language**

- *Every cat is on a mat*
- *John is going to Boston by bus*
- *Tom believes that Mary wants to marry a sailor*

http://www.jfsowa.com/cg/cgexamp.htm



## FrameNet

Relationships among conceptual frames:
Is-a, subframe, perspective, inchoative, causative, precedence, etc.

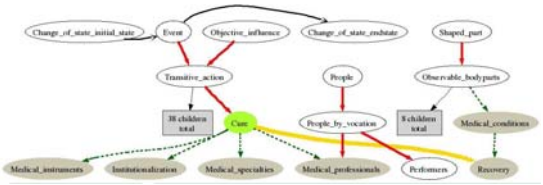



Image obtained with http://framenet.icsi.berkeley.edu/FrameGrapher/grapher.php



## Online access to Spanish FrameNet




## "Word" in FrameNet

- When we say that the word *bake is* polysemous, we mean that the lemma *bake.v* (which has the word-forms *bake, bakes, baked,* and *baking*) is linked to three different frames:
  - Apply heat: Michelle baked the potatoes for 45 minutes.
  - Cooking creation: Michelle baked her mother a cake for her birthday.
  - Absorb heat: The potatoes have to bake for more than 30 minutes.
- These constitute three different Lus [lexical units], with different definitions.
- Multiword expressions such as *given name* and hyphenated words like *shut-eye* can also be LUs.

Ruppenhofer et al. (2010)



## What are word senses?

- Is word sense disambiguation just one more NLP task? (Wilks 2000)
- *Hot*, *warm* and *cold* (Ellis, 1993)
  - Particular and arbitrary ranges of temperatures are associated with these words
  - Not different in kind from measurements, simply a very primitive system of measurement
  - Every language is a particular system of classification
- Cruse (2004) on several criteria



## Back to the beginning?

- From a Web advertisement
- *toy for generating "word clouds" from text that you provide. The clouds give greater prominence to* ***words that appear more frequently*** *in the source text. You can tweak your clouds with different fonts, layouts, and color schemes.*


http://www.readwriteweb.com/archives/tag_clouds_of_obamas_inaugural_speech_compared_to_bushs.php

## Concluding remarks

- There is a huge activity nowadays in (automatically or not) creating (lexical or not)) ontologies and merging or integrating them
- Unfortunately, many of the work is still based on ungrounded or naive assumptions
  - What is similarity
  - What is the purpose of the O
  - What are its units
- There are a lot of fancy tools and systems to deal with and visualize complex objects created from heaps of data

  **but their use is only as good as the underlying objects...**

WATS:: Dictionaries, lexical networks, lexical ontologies, wordnets and wordclouds 37

## Lexical statistics

Diana Santos

**WATS: Words and their secrets**, ESSLLI 2010
Diana Santos & Maria José Bocorny Finatto

Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil



## Preview

- What is statistics?
- Why do people use statistics in linguistics? (good and bad reasons)
- Why do people use linguistics in statistics? (good and bad reasons)
- Lexical statistics:
  - comparing the meaning of words
  - words representative of text
  - modelling the occurrence of words
- Zipf's law, Mandelbrot's law, long tail...
- Typical applications
  - MWE's revisited
  - Machine translation and BLEU
  - Indexing



## Statistics is the branch of mathematics...

- That is concerned with uncertainty ☑
- That is most frequently used in non-hard-sciences, that is: medicine, sociology, literature ☑
- That is harder to teach in schools ☐
- That is most used in real life applications ☑
- That is most abused/misused in newspapers and political speeches ☐
- That is less understood by practictioners of nearby sciences (and this includes language sciences ☺) ☑
- On which there are more dedicated textbooks ☐



## Why use statistics in linguistics...

- To account for lack of sufficient information, ... or due to the probabilistic nature of the information available (Katz, 1996)
- Halliday (2005): "probability" as a theoretical construct is just the technicalising of "modality" from everyday grammar
  - The grammar of a natural language is characterized by overall quantitative tendencies (two kinds of systems)
    - equiprobable: 0.5-0.5
    - skewed: 0.1-0.9 (0.5 redundancy) – unmarked categories
  - In any given context, ... global probabilities may be significantly perturbed. ... the local probabilities, for a given situation type, may differ significantly from the global ones. *"resetting" of probabilities ... characterizes functional (register) variation in language*. This is how people recognize the "context of situation" in text. (pp. 236-8)



## Why use statistics in linguistics...

- *The field of statistical NLP is very young, and the foundations are still being laid*... Magerman, 1995, apud Krenn & Samuelsson (1997):
- "The stability of the relative frequency": there is some structure even in random processes... The relative frequency stabilizes around a number after a large number of trials --> this number is its probability
- To guide us in the maze of **LARGE** amounts of data
- Because we want
  - to have independent criteria for choice
  - to have independent criteria for sampling
  - to have independent criteria for evaluation



## Why use linguistics in statistics

- Because "everyone" knows words and basic grammar
- Because there is a third discipline which is connected to both: information theory, and coding/cyphering/criptology
- I have an answer! I have an answer! Are there any questions around? ☺

## Why connect the two disciplines at all?

- Because speech processing – itself inspired by probability theory – has been influential as a model for empirical language processing
  - It remains to be assessed how successful statistical methods for speech have actually been
  - It remains to be assessed how close speech processing in fact is to machine translation (for example)
- Because information retrieval – itself making heavy use of probabilistic models – has also been influential as a testbed for empirical language processing



## Lexical statistics are concerned with words

- Apparently this does not help us much, because we know **words** may even be a more elusive concept than statistical ones...
- Statistics are associated with large numbers; and therefore with computers
  - Simulation
  - Random sampling
  - Multivariate techniques
  - Reduction techniques
- See Baayen (2001) for the standard reference on this subject
- Statistics is concerned with counting... filtering ... helping to organize



## The dispersion index (Calzolari & Bindi, 1990)

- Measures the degree of fixity of the second word position with respect to the keyword, a measure of how frequency is distributed over the different positions of the window...
- Different slices of the multidimensional pie (the semantic hyper-space) carry with themselves a different bunch of word senses for the same word entry (Bindi et al., 1994)

- Italian quasi-synonyms: *picollo*, *corto*, *breve*, *ristretto*, *esiguo*, *scarso*, *ridotto* are the "units"
- Space: "word mates" (unambiguous words with whom the units keep company with, computed by mutual information)



## The overall representation (Bindi et al. 1994)

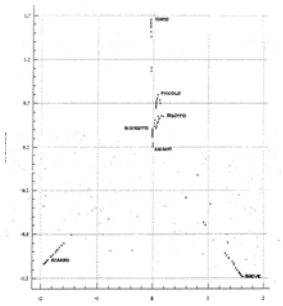


Fig. 3 from Bindi et al (1994:42)



## Which words are particularly characteristic of a text? (Kilgarriff, 1996)

Survey of different statistical approaches

- $\chi^2$ test
- Mann-Whitney ranks test
- *t*-test
- Mutual information (MI)
- Log-likelihood ($G^2$)
- tf-idf (term frequency, inverse document frequency)
- Poisson mixtures: "Documents are more than just a bag of words" CG
  - Adjusting frequencies to reflect clumpiness
- Multi-dimensional analysis



## Correlation between IDF and word frequency

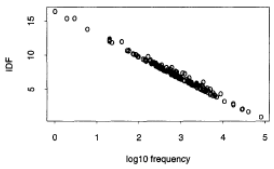



Figure 1: IDF is highly correlated with log frequency ($\rho = -0.994$). The circles show $\log_{10} f$ and IDF for 193 words selected from a corpus of 1989 Associated Press Newswire stories ($D = 85{,}432$).

Church & Gale (1995a:122)

## Predicting the occurrence of words

- A good keyword is one that behaves very differently from the null hypothesis (that the word is distributed according to a Poisson distr.)
- Variance and IDF correlate positively with good keywordness, entropy negatively
- Katz K-mixture has two parameters (α: fraction of relevant and irrelevant documents, and β: the average Poisson parameter) and corresponds to a convolution of Poisson distributions
  - $\beta = f/D*2^{IDF}-1$
  - $\alpha = f/D\beta$

Church & Gale (1995a)

- The main idea is that each Poisson distr. can model hidden variables such as *what the documents are about, who wrote them, when they were written, what was going on in the world then*



## Clumpiness, burstiness and other properties

- *Content words like* Kennedy *tend to be very contagious*
- *Text is more like a contagious disease than lightning*
- Measures of variability, and their empirical estimates
  - Variance
  - Entropy
  - Burstiness
  - Adaptation
  - ...

$$E_{NB}(k) = NP \quad \text{mean}$$

$$\sigma^2_{NB} = NPQ \quad \text{variance}$$

$$IDF_{NB} = -\log_2 Pr_{NB}(x \geq 1) = -\log_2(1-Q^{-N}) \quad \text{inverse document frequency (IDF)}$$

$$H_{NB} = -\sum_{k=0}^{\infty} Pr_{NB}(k)\log_2 Pr_{NB}(k) \quad \text{entropy (H)}$$

$$B_{NB} = \frac{NP}{1-Q^{-N}} \quad \text{burstiness (B)}$$

$$Pr_{NB}(k \geq 2 | k \geq 1) = \frac{Pr_{NB}(k \geq 2)}{Pr_{NB}(k \geq 1)} = \frac{1-Q^{-N}-NPQ^{-N-1}}{1-Q^{-N}} \quad \text{adaptation}$$

- *There ought to be a quantity discount*

Church & Gale (1995b:170)



## "Stopwords": uninteresting words

- Mosteller & Wallace (1964) put content words in stoplists
- IR in general puts grammatical words in stoplists

- Different distribution in a same text and in a collection of texts: between and within documents
  - Different distribution in different genres
  - Different distribution in different authors
  - Different distribution in different themes
  - ...



## Katz's (1996) model of distribution of words in text

- Starting with the texts themselves and the way they come about
- The main players are the content words (which define the function words they require), and their number and repetition is dependent on the message to convey.
  - The frequency of function words (in large enough documents) is proportional to the document length
  - The frequency of content words depends on their **topicality**, and only related to document length indirectly
  - When a content word is topical, it displays **multiple** and often **bursty** occurrence (so, a word can be unrelated to a document, non-topical, and topical, and this shows in 0,1, >1 occurrences in the document)
  - Two kinds of burstiness: document-level, and within-document
- Length of a text is the number of occurrences of *the* (instead of blanks)



## Katz (1996) continued

- Linguistically motivated approach (...) arriving at a coherent view of the word occurrence phenomenon without commitments to any particular, *a priori* assumed, stochastic mechanism
  - The probabilities of repeat occurrences do not depend on the relative frequency
  - The continual presence of repeat occurrences in discourse is a general and widespread phenomenon (...) A principle distinction is identified between two probabilities of repeats (entering; and stayin in a document-level burst)
- Possion mixtures as two-stage stochastic mechanism for generating content words is incompatible with empirical data
- (discrete Poisson mixtures) limited in their capacity to provide satisfactory fit to the data because of their faulty functional form...



## Green (1979) and syntax markers

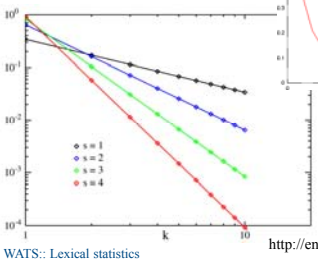
- The marker hypothesis states roughly that "a small number of elements that signal the presence of particular syntactic constructions" is required in order for a language to be learnable
- Markers in English: prepositions/closed words, suffixes such as *-ly* or *–ing*

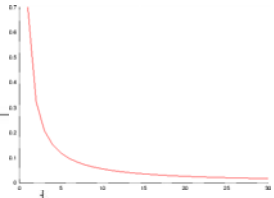
- This is interesting food for thought also for natural language, although the issue of what markers are is obviously subject to the same kind of problems about words (what are the units?)

## Zipf's law

http://planetmath.org/encyclopedia/ZipfsLaw.html

- Rank and frequency are inversely proportional



$$O(i) = \frac{n}{i^{\theta} H_{\theta}(V)}$$

http://en.wikipedia.org/wiki/Zipf%27s_law



## Zipf's law: discussion

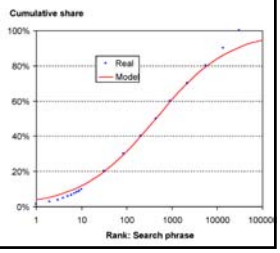
- A law in which sense? A statistical regularity...
- Is this related no natural language, or to mankind in general?
  - Zipf suggested it for human avoidance of work (in all respects)
  - Mandelbrot developed fractals inspired by it
- Are there different coefficients
  - Per language?
  - Per objects? (forms, lemmas, grammatical categories, etc.)
- How relevant is it at all?
  - Is it also true of randomly generated "texts"?



## The long tail (Kilkki, 2007)

$$F(x) = \frac{\beta}{\left(\frac{N_{50}}{x}\right)^{\alpha} + 1}$$

- In essence, the phrase "long tail" refers to those numerous objects that have very limited popularity but that together form a significant share of the total volume
- $N_{50}$ is the share of the objects that cover half of the whole volume
- $\beta$ total volume; $x$ is the rank






## Two critics of "typical" SLP (statistical language processing)

- Dunning (1993)
  - Mutual information, with the assumption of maximum likelihood estimate to estimate probabilities from frequencies, fares poorly when estimating the probabilities of rare events – which are the vast majority of interesting events in linguistics
- Kilgarriff (2005)
  - The probability model, because of its assumption of randomness, is inappropriate for large numbers.
  - The null hypothesis is never true... because language is not random
  - Instead of testing the null hypothesis, they are merely testing whether they had enough data to reject the null hypothesis with confidence...



## Multi-word expressions

- Church and Hanks (1991) proposed a word association measure ... to *help lexicographers organize a concordance*.
- Justeson & Katz (1995) looked at the distribution of terminology in text, proposing
  - frequency features in an in-document characterization of terminology
  - structural features of the terms themselves



## Machine translation and its evaluation

**BLEU (Papineni et al, 2001)**

- using *n*-gram similarity of a candidate to a set of reference translations (sentence based)
- modified precision:

  number of clipped words (n-grams) that occur in the candidate / number of total words (n-grams) in the candidate

  sum of clipped n-grams in all sentences / sum of candidate n-grams
- *word-weighted average of sentence-level modified precisions, rather than a sentence-weight average*
- combination of the modified precisions of 1 to 4 grams
- sentence-brevity penalty

## Example from Papinemi et al (2001)

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

P1=17/18
P2=5/18

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

WATS:: Lexical statistics 25

## BLEU formulas

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad \text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n$$

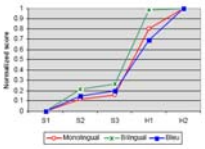
- c, r – length of the candidate or reference translations
- As a baseline, Papineni et al suggest:
  - wn – uniform weights: 1/N
  - N = 4
- Note that the matches are position independent.

WATS:: Lexical statistics 26

## More on BLEU (Bilingual Evaluation Understudy)

- Proposed for use in the R&D cycle of machine translation technology
- The more reference translations, the higher the precision
- Even a human translator will hardly score 1 (except if s/he produces a translation equal to one of the reference translations)
- experiments to judge 5 "systems":
  - 250 Chinese-English sentence pairs
  - rated by two groups of human judges
  - from 1 (very bad) to 5 (very good)
  - 10 bilinguals and 10 monolinguals
  - 5 translations of each sentence
  - linearly normalized by the range

Figure 7: BLEU vs Bilingual and Monolingual Judgments




WATS:: Lexical statistics 27

## Indexing

- This is the realm of information retrieval...
- Or the use of good "descriptors": what best than words themselves?
- Sparck Jones (2004) "lessons from information retrieval"
  - away from lexical normalisation and towards relational simplification
  - decreasing ontological expressiveness, epistemological commitment, and inferential power
  - Shallow text operations (...) are right for information access. Information is primarily conveyed by natural language and this has to be shown to the user for them to assess
- and Wilks & Brewster (2009) state: The Semantic Web is nothing else other than scaling up natural language processing...

WATS:: Lexical statistics 28

## Vagueness, ambiguity and multilingual issues

Diana Santos

**WATS: Words and their secrets**, ESSLLI 2010
Diana Santos & Maria José Bocorny Finatto

Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil



## Preview

- Vagueness, a key concept
- Contrastive studies: three models according to Santos (1996)
- Contrastive studies: four models according to Pinker (2007)
- Contrastive studies: three models according to Chesterman (1998)
- Several formalizations of translation and constrasts
  - Catford notion of functional relevance
  - Snell-Hornby discussion of "descriptive verbs"
  - The translation network
- Clines involving words: Ellis, Halliday, Talmy, etc.
- *To be or not to be ...*



## Properties that define a natural language as opposed to artificial ones

1. Metaphorical nature
2. Context dependency
3. Reference to implicit knowledge
4. **Vagueness**
5. Dynamic character (evolution and learnability)

Slide 12 from Santos (2006)



## 4. Vagueness: the most important property

- The same unit means more than one related thing, at the same time.
- Crucially different from ambiguity:
  - although both give more than one translation to one entity
  - the difference is in the **relationship** among the translations
  - vagueness is **systematic**, ambiguity is accidental
- Vagueness has been the subject of much linguistic-philosophical research (Quine, Dahl, Lakoff, Kempson, Lyons, Keenan, etc. etc.) but it is somehow considered a nuisance for NLP

Santos, Diana. "The relevance of vagueness for translation: Examples from English to Portuguese". *TradTerm* Vol. 5.1, 1998, pp. 41-98.

Slide 32 from Santos (2006)



## Vagueness, polysemy and underspecification

- Vagueness is the general property: positively meaning related things
- Polysemy is vagueness restricted to the lexicon (related word senses)
- Underspecification is a more general name that includes vagueness: one might say that e.g. *table* is unspecified wrt weather, but not vague about the weather
- Vagueness is essential for communication, learning and evolution...

Slide 36 from Santos (2006)



## Selected examples of vagueness

- *Apaixonado, recusou o convite*
- It can be translated by: "in love, he refused", or "of a passionate character, he refused"
- *Encontraram-se na praia*
- Can be translated by "They met on the beach" or "They found themselves on the beach"
- *A porta abriu-se!*
- Can be translated as "Someone open the door" or "The door opened (itself)"
- *The man who killed X is mad!* (atribution, or description?)
- or, and: inclusive or exclusive; causal or logical?

## Vagueness at all levels of description

- POS: the infamous case of past participles
- The case of *near*: adjective or preposition? (Manning & Schütze, 1999)
- The most famous case is, however, PP attachment. After discarding non V NP PP structures, Hindle and Rooth state:
- *Disambiguating the test sample turned out to be a surprisingly difficult task. [...] more than 10% of the sentences seemed problematic to at least one author* (Hindle & Rooth, 1993:112)



## Attempts to deal with vagueness

- In annotation, leave room for more than one category: HAREM and COMPARA
  - do not force a choice when it is not required
- Identify contrastively vague categories in tense and aspect
  - not only coercion
  - also aspectual classes or grammatical operators that can simultaneously mean more than one thing
- The translation network
  - linking two systems with different vague categories
  - explaining and formalizing concrete translation issues

Slide 49 from Santos (2006)



## Again: How does a language choose its units?

- Talmy's (1983:277ff) suggestion:
- *The majority of semantic domains in language are n-dimensional, with n a very large number. For example, no fewer that [] twenty parameters are relevant to the domain of spatial configuration as expressed by closed-class elements such as English prepositions and deictics. [List]*
- *With so many parameters, full domain coverage by fairly specific references would require thousands of distinct vocabulary items, […]*
- *Rather that a contiguous array of specific references, languages instead exhibit a* ***smaller number*** *of such references in a* ***scattered distribution*** *over a semantic domain. That is, a fairly specific reference generally does not have any immediate neighbors of equal specificity.*



## Cont.

- *General terms are necesssary for referring to insterstitial conceptual material, between the references of specific terms*
- *Their locations must nevertheless be to a great extent arbitrary, constrained primarily by the requirement of being "representative" of the lay of the semantic landscape, as evidenced by the* ***enormous extent of non-correspondence between specific morphemes of different languages****, even where these are spoken by the peoples of similar cultures.*

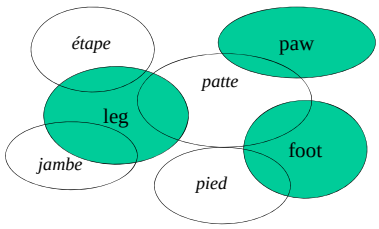


## Examples

- Bowerman (1996), child language acquisition of Korean and English
- Pinker (2007), spatial reasoning
- Sampson (2005/1997), interesting distinctions
- Dixon (1971) and Dyirbal's "mother-in-law language"
- Santos (1996), choice of permanent or temporary property
- Numbers:
  - how many lives does a cat have?
  - How many heavens are there?
  - How many days there is in a fortnight?
  - How many divisions there is in a clock?



## Jurafsky & Martin (2000:806) lexical overlap



## Contrastive studies (according to Santos 1996)

- Universalism
  assume that differences are noise, and that they can be parametrized and done away at a deep enough level)
- Typology
  classify all languages on a number of axes, on the search of universal or frequent traits
- Relativism
  take all languages as equals: the only unbiased way



## Contrastive studies (according to Pinker 2007)

Theories of language, in Pinker's (2007) words

- Extreme Nativism: *born with 50,000 concepts (Fodor)*
- Radical pragmatics: *people can use a word to mean almost anything (Sperber and Wilson)*
- Linguistic determinism: *words determine thoughts (Sapir and Whorf)*
- Pinker's moderate position ☺: *meanings of words are formulas in an abstract language of thought*



## Contrastive studies (according to Chesterman 1998)

Overview of the concept of equivalence in Translation Theory (pp.16-27)

- The equative view
  *Signs represent meanings; meanings are absolute, unchanging, they are manifestations of the ideal, they are Platonic Ideas*
  *identity of meaning across translation*
- The taxonomic view
  *Different types of equivalence are argued to be appropriate in the translation of different kinds of texts*
  *Nida's formal equivalence vs dynamic equivalence*
- The relativist view



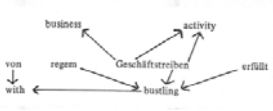
## Three ways of arriving at the relativist view

- From rational thinking: Logical rejection of sameness, replacing it by similarity, matching or family resemblance, or economical considerations
- *equivalence depends only on what is offered, negotiated and accepted in the exchange situation* (Pym, 1992/2010:46)
- From cognition: *the interpretation of an utterance is a function of the utterance itself and the cognitive state of the interpreter: we interpret things in the light of what we already know...*
- From comparative literature and translation: *TS is an empirical science whose aim is to determine the general laws of translation behaviour. Translations have many purposes and are of many kinds*



## Trying to make sense of language differences

- How to indicate the relationship of meaning "nuggets" in different languages?
- Snell-Hornby (1983) on the translation of German *von regem Geschäftsstreibem erfüllt*






## Snell-Hornby's descriptive verbs by semantic area

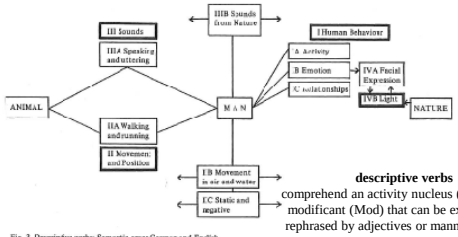
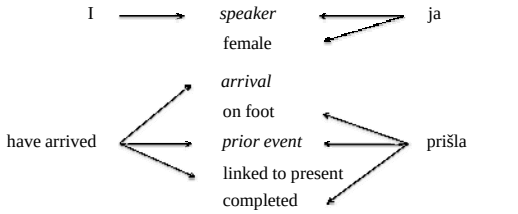



Fig. 3 Descriptive verbs: Semantic areas German and English

**descriptive verbs**
comprehend an activity nucleus (ANu) and a modificant (Mod) that can be expressed or rephrased by adjectives or manner adverbs, and often carries speaker's evaluation on some of the agents or on the action itself (Snell-Hornby, 1983)

## How can two sentences be translations of each other?

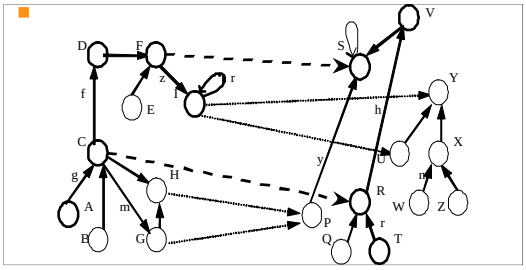
… a SL and a TL text or item [being] relatable to (at least some of) the same features of substance (Catford, 1967:50)




Catford (1967:39)



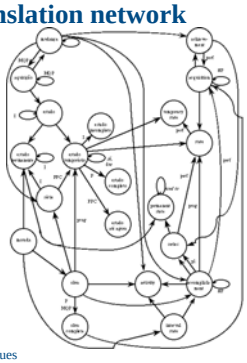
## A translation pair in a translation network





## Portuguese-English translation network

- Using a formalization of two languages' tense and aspect systems and observing the translation from one and into the other
- And the other way around, a different E-P TN





## Steiner's *mystère supreme* of anthropology

- Why does *homo sapiens* whose digestive tract has evolved and functions in precisely the same complicated ways the world over, ... -- why does this unified, though individually unique mammalian species does not use one common language? (Steiner, 1992 [1975] :52)



## *To be* or not to be: that's the question?

- It is remarkable how the verb *to be* is a complex problem for linguitsic description, and for translation, whose interpretation of this famous quote is difficult, to say the least
- *The interpretation of* ***be*** *is an interesting chapter of natural language semantics. For the present purpose, it is enough to say that it ambiguously represents the operations of identity, membership and class inclusion.* (Carlson, 1981: 156)
- *the ambiguous noun* time (Carlson, 1981:60) is translationally vindicated in Portuguese as follows: as a count noun, *time* is translated in Portuguese by *vez* ("turn"); as a mass noun, it represents the temporal domain (*tempo*). Cf. no. *gang* ("going"), fr. *fois*, it. *volta* …



## Concluding remarks

- It is hardly to be found ONE distinction that is common across all natural languages
- Languages tend to evolve and age and innovate continuously
- The comparison of languages is arguably the best mirror into language

... and the comparison itself is best done through translation data

- Words carve different domains in different languages, words are different in different languages, the differences between inter-translatable words (and not only) are a wonderful mirror to differences in systematic organization of the languages (systematicy includes creativity)

## Words and their secrets: Conclusion(s)

Diana Santos & Maria José Bocorny Finatto
ESSLLI 2010



## What have we tried to teach?

- Monday
  - Introduction (D+MJ)
  - Linguistic evolution: from words in the mind to real utterances (MJ)
- Tuesday:
  - Basic technologies: spell checking and POS tagging (D)
  - Word types and their function in texts (MJ)
- Wednesday
  - Dictionaries, lexical networks, lexical ontologies, wordnets and wordclouds (D)
  - Lexicography and terminography: old traditions and new routes (MJ)
- Thursday
  - Frequency studies in Portuguese: *de* and *Brasil* (MJ)
  - Lexical statistics (D)
- Friday
  - Vagueness, ambiguity, and multilingual issues (D)
  - Conclusions (MJ+D)



## On Monday

- Scare you: Beware that words are not that simple!
- There are many many issues related to the concept of word
- There have been many different answers throughout history...



## On Tuesday

- The simplest NLP applications are not that simple after all
  - Tokenization
  - Spell checking
  - PoS tagging
- and they depend crucially on the notion of what a word is: also, the momentous issue of types versus tokens



## On Wednesday

- MJ: terminology: terms versus words?
- D: An overview of several methods of representing the collection of words in one language
- Again, many assumptions and choices that we tried to highlight, and which require a clear notion of word properties
- And the type/token distinction oneness/individuation reopened



## On Thursday

- D: Some statistical tools to investigate words .. and how many assumptions are required again
- From simple counting to making sense of counts at all
- MJ: A detailed example

## Today

- D: vagueness: THE property of natural language
- Words are different in different languages: from this fact to the many possible inferences that can be brought to bear on this
- And the notion of word illuminated as well

- Wrap up

SINTEF Information and Communication Technologies
Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil UFRGS 7

## We would like some evaluation

- What were you expecting that was not dealt with?
- What was it that was too easy – or too difficult?
- What you would not have here but we brought anyway?

- If a further / advanced course on WATS were to be prepared, which areas would you like to see covered?
- Would you attend it?

- Thank you for your participation!

SINTEF Information and Communication Technologies
Institute of Language and Linguistics, Federal University of Rio Grande do Sul, Brazil UFRGS 8