Translation and Tag-based Search

Shibamouli Lahiri

(shibamouli@cse.psu.edu)

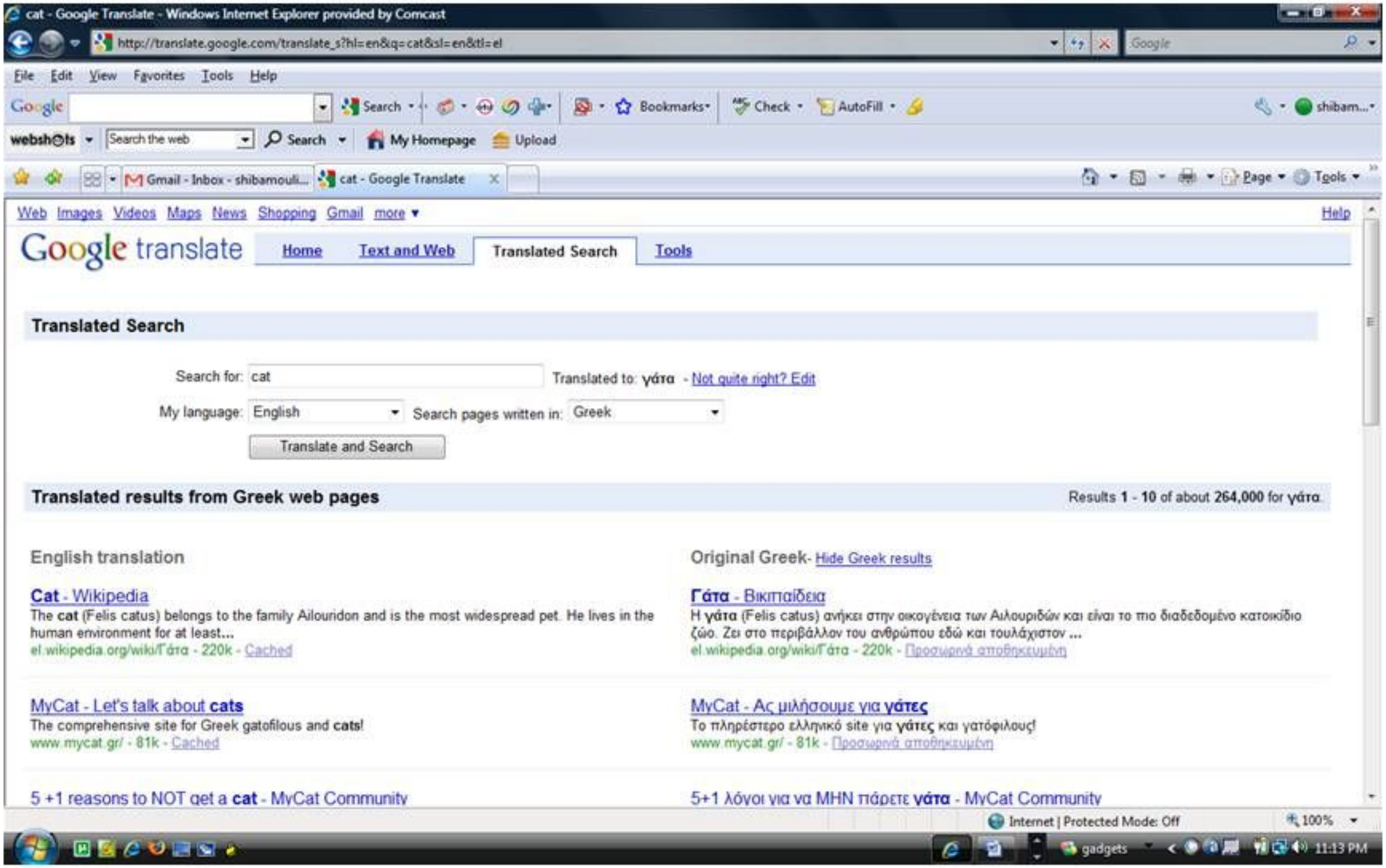
Outline

- **Brief recap**
- Translation issues
- How can we use tags?

A Short Recap

- Search Engine on Ancient Greek corpus
- Query in Ancient Greek, Modern Greek or English

A Short Recap



Issues Identified

- Translation/transliteration
  - ➢ Disambiguation[3,4,5,8]
  - ➢ Evaluation[9,10,11]
- Unit of Retrieval (pages or paragraphs)
- Query disambiguation
- Ranking
- Retrieval evaluation
- Named Entity Extraction[6]
- Summary in three languages

Outline

- Brief recap
- **Translation issues**
- How can we use tags?

Why Google Translate does not help?

- Herodotus, "The Histories", Book 1, Chapter 1, Section 0

- Ancient Greek: "Ἡροδότου Ἁλικαρνησσέος ἱστορίης ἀπόδεξις ἥδε, ὡς μήτε τὰ γενόμενα ἐξ ἀνθρώπων τῷ χρόνῳ ἐξίτηλα γένηται, μήτε ἔργα μεγάλα τε καὶ θωμαστά, τὰ μὲν Ἕλλησι τὰ δὲ βαρβάροισι ἀποδεχθέντα, ἀκλεᾶ γένηται, τά τε ἄλλα καὶ δι᾽ ἣν αἰτίην ἐπολέμησαν ἀλλήλοισι."

- Human translation (A.D. Godley, Harvard University Press, 1920): "This is the display of the inquiry of Herodotus of Halicarnassus, so that things done by man not be forgotten in time, and that great and marvelous deeds, some displayed by the Hellenes, some by the barbarians, not lose their glory, including among others what was the cause of their waging war on each other."

- Google Translate: "Herodotus Alikarnisseos istoriis apodexis ide, ὡς nor τὰ contemplated ἐξ people unto χρόνῳ genitai faded, nor works great II. THOMAS, as regards the ELLIS And the varvaroisi were accepted, aklea genitai, GFs, but also self aitiin epolemisan not they,."

- Yahoo Translate (AltaVista): "H[rodotoy] A[likarnisseos] i[storiis] a[podexis] h[de], w[s] [mite] [t]a [genomena] e[x] a[nthropon] [t]w [chron]w e[xitila] [genitai], [mite] e[rga] big [te] [ka]i [thomasta], [t]a [m]e[n] E[llisi] [t]a [d]e [barbaroisi] a[podechthenta], a[kle]a [genitai], the [te] a[lla] [ka]i [di]' h[n] [a]i[tiin] e[polemisan] a[lliloisi]."

Parallel Corpus

- Perseus Project (Ancient Greek ←→ English)
- 87 books by 30 authors (4M Greek words and 6M English words)
- Europarl[1] has almost 30M for each language
- Koehn and Monz[2] used 15M training words for translation ({Spanish, French, Finnish, German} → English) and roughly 60K test words
- Literary Translation → alignment issue[12,13]

Open Source

- Giza++
- Moses

Other Corpora

- Project Gutenberg
- Greek Bible
- Sacred Texts

Ancient Greek Dictionary

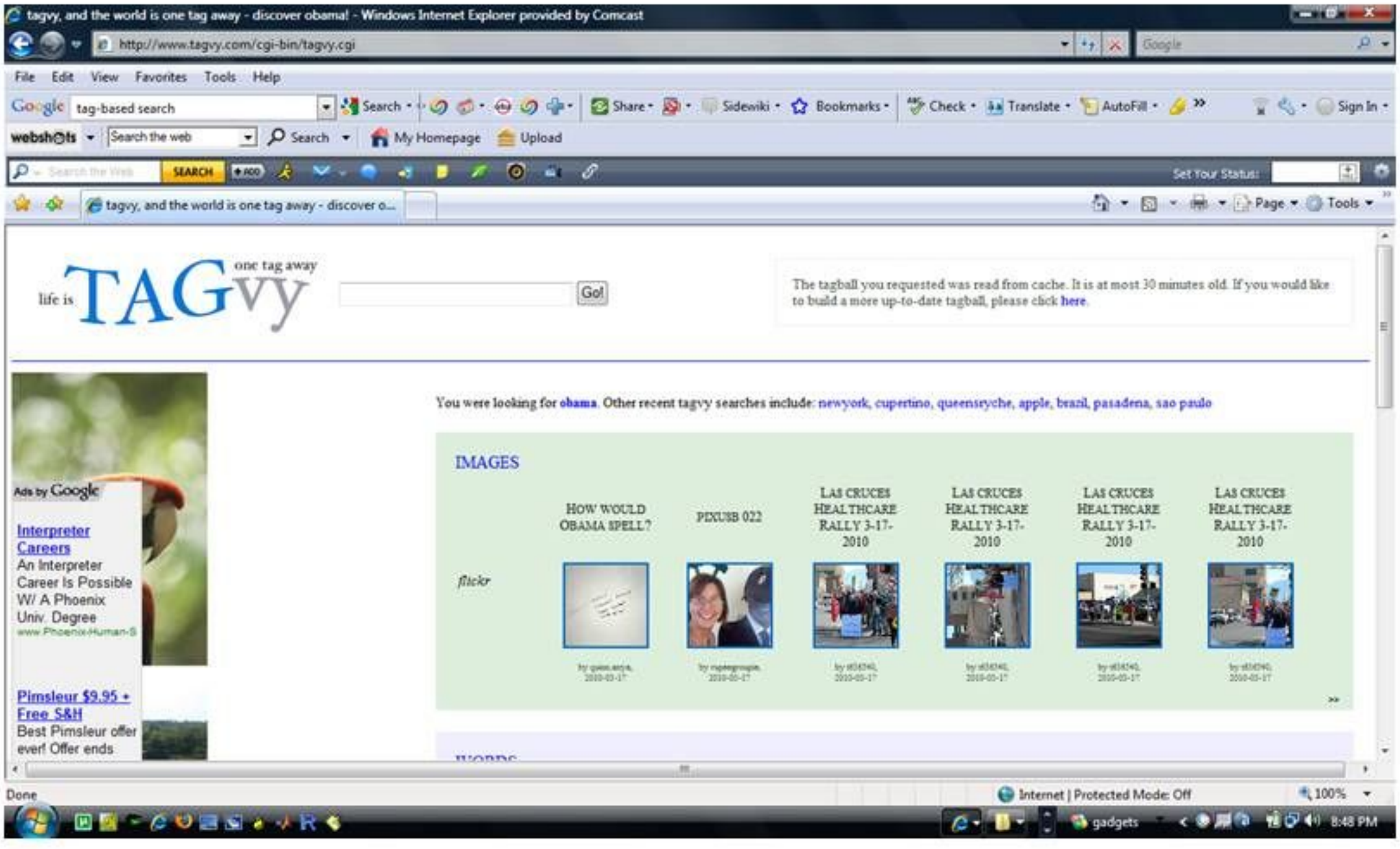
- Lexilogos
- Kypros
- Translatum
- Ectaco

Outline

- Brief recap
- Translation issues
- **How can we use tags?**

Tag-based Search

Tag-based Search



How can tags help?

- Useful information regarding an object of interest
- Blogs, microblogs and multimedia search
- Tag clouds – the larger the tag, the higher its use
- Make "semantic" queries possible
- Query expansion[14] and Disambiguation[15]

Issues

- Language of tags (Greek or English)
- Level of tags (pages, paragraphs or sentences)
- Tags have their own set of problems[7]
  - Coverage
  - Ambiguity
  - Noise
  - Spam tags

References

1. **Europarl: A Parallel Corpus for Statistical Machine Translation.** Philipp Koehn. *Machine Translation Summit (2005).*
2. **Shared Task: Statistical Machine Translation between European Languages.** Philipp Koehn and Christof Monz. *Proceedings of the ACL Workshop on Building and Using Parallel Texts (June 2005).*
3. **Using Structured Queries for Disambiguation in Cross-Language Information Retrieval.** David A. Hull. *AAAI (1997).*
4. **Using the Web for Translation Disambiguation.** Y. Zhang and P. Vines. *NTCIR-5 Workshop (2005).*
5. **Query Disambiguation for Cross-Language Information Retrieval Using Web Directories.** F. Kimura, A. Maeda, J. Miyazaki and S. Uemura. *Web Information Retrieval and Integration (2005).*
6. **Proper name translation in cross-language information retrieval.** H.H. Chen, S.J. Huang, Y.W. Ding and S.C. Tsai. *ACL (1998).*
7. **Information Seeking with Social Signals: Anatomy of a Social Tag-based Exploratory Search Browser.** Ed H. Chi and Rowan Nairn. *International Conference on Intelligent User Interfaces (2010).*

References (Contd.)

8. **Iterative Translation Disambiguation for Cross-Language Information Retrieval.** Christof Monz and Bonnie J. Dorr. *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2005).*
9. **BLEU: a Method for Automatic Evaluation of Machine Translation.** K. Papineni. *ACL (2002).*
10. **A paraphrase-based approach to machine translation evaluation.** G. Russo-Lassner, J. Lin and P. Resnik. *(2005)*
11. **ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation.** Chin-Yew Lin and Franz Josef Och. *ACL (2004).*
12. **Text-Translation Alignment.** M. Kay and M. Röscheisen. *Computational Linguistics (1994).*
13. **HMM-based word alignment in statistical translation.** S. Vogel, H. Ney and C. Tillmann. *ACL (1996).*
14. **Query expansion using lexical-semantic relations.** E. M. Voorhees. *SIGIR (1994).*
15. **Senseval: An exercise in evaluating word sense disambiguation programs.** A. Kilgarriff. *Proceedings of LREC (1998).*