Morphology and Corpora: Introduction

Marco Baroni

University of Bologna

Granada “Morphology and Corpora” Seminar

Outline

Corpora

- General overview
- Data sparseness and the need for larger corpora

Morphology

- Derivational vs. inflectional morphology
- Data in morphology

Corpora: what and why

- Collections of natural text stored on computer

Corpora: what and why

- Collections of natural text stored on computer
- Useful for:

Corpora: what and why

- Collections of natural text stored on computer
- Useful for:
  - NLP (e.g., speech recognition, text categorization, question answering, machine translation. . .)

Corpora: what and why

- Collections of natural text stored on computer
- Useful for:
  - NLP (e.g., speech recognition, text categorization, question answering, machine translation. . .)
  - lexicography, grammar writing, language teaching

Corpora: what and why

- Collections of natural text stored on computer
- Useful for:
  - NLP (e.g., speech recognition, text categorization, question answering, machine translation. . .)
  - lexicography, grammar writing, language teaching
  - theoretical linguistics?

Typology

- Balanced, representative, ‘reference’ corpora: Brown/LOB (1M tokens), COBUILD (10M, . . .), BNC (100M)

Typology

- Balanced, representative, ‘reference’ corpora: Brown/LOB (1M tokens), COBUILD (10M, . . .), BNC (100M)
- Opportunistic: WSJ, la Repubblica-SSLMIT, Gigaword (1B)

Typology

- Balanced, representative, ‘reference’ corpora: Brown/LOB (1M tokens), COBUILD (10M, . . .), BNC (100M)
- Opportunistic: WSJ, la Repubblica-SSLMIT, Gigaword (1B)
- Web-derived corpora (WaCky project: 1.65B tokens of German, 1.9B tokens of Italian)

Typology

- Balanced, representative, ‘reference’ corpora: Brown/LOB (1M tokens), COBUILD (10M, . . .), BNC (100M)
- Opportunistic: WSJ, la Repubblica-SSLMIT, Gigaword (1B)
- Web-derived corpora (WaCky project: 1.65B tokens of German, 1.9B tokens of Italian)
- Specialized, parallel, comparable, diachronic. . .

Standard requirements for a modern corpus

- POS-tagging and lemmatization

Standard requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing with specialized software that allows sophisticated linguistic queries

Standard requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing with specialized software that allows sophisticated linguistic queries
- Many other desirable features:

Standard requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing with specialized software that allows sophisticated linguistic queries
- Many other desirable features:
  - Meta-data

Standard requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing with specialized software that allows sophisticated linguistic queries
- Many other desirable features:
  - Meta-data
  - Syntactic parsing

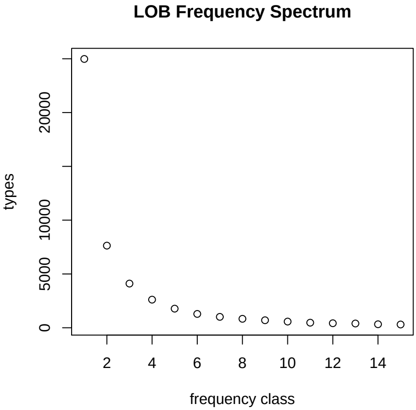
Standard requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing with specialized software that allows sophisticated linguistic queries
- Many other desirable features:
  - Meta-data
  - Syntactic parsing
  - Web interface

Standard requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing with specialized software that allows sophisticated linguistic queries
- Many other desirable features:
  - Meta-data
  - Syntactic parsing
  - Web interface
  - …

Zipf's Law


LOB Frequency Spectrum
types
0
5000
10000
20000
2
4
6
8
10
12
14
frequency class

There is no data like more data!

- In NLP (Banko and Brill, 2001), lexicography (Kilgarriff 2005) as well as corpus-based linguistics (Mair, 2003), often…

There is no data like more data!

- In NLP (Banko and Brill, 2001), lexicography (Kilgarriff 2005) as well as corpus-based linguistics (Mair, 2003), often…
- more data is better data!

There is no data like more data!

- In NLP (Banko and Brill, 2001), lexicography (Kilgarriff 2005) as well as corpus-based linguistics (Mair, 2003), often…
- more data is better data!
- This implies:

There is no data like more data!

- In NLP (Banko and Brill, 2001), lexicography (Kilgarriff 2005) as well as corpus-based linguistics (Mair, 2003), often…
- more data is better data!
- This implies:
    - Less clean data sources (the Web)

There is no data like more data!

- In NLP (Banko and Brill, 2001), lexicography (Kilgarriff 2005) as well as corpus-based linguistics (Mair, 2003), often...
- more data is better data!
- This implies:
  - Less clean data sources (the Web)
  - Automated processing

Outline

Corpora

- General overview
- Data sparseness and the need for larger corpora

Morphology

- Derivational vs. inflectional morphology
- Data in morphology

Derivation vs. inflection

- Derivational morphology: word formation, e.g.: compounding, nominalizations, English prefixing

Derivation vs. inflection

- Derivational morphology: word formation, e.g.: compounding, nominalizations, English prefixing
- Inflectional morphology: syntax-driven morphology, e.g.: agreement, plural formation, verbal paradigms

Derivation vs. inflection

- Derivational morphology: word formation, e.g.: compounding, nominalizations, English prefixing
- Inflectional morphology: syntax-driven morphology, e.g.: agreement, plural formation, verbal paradigms

Derivation vs. inflection

- Derivational morphology: word formation, e.g.: compounding, nominalizations, English prefixing
- Inflectional morphology: syntax-driven morphology, e.g.: agreement, plural formation, verbal paradigms
- Corpus data especially relevant to derivational morphology (productivity, lexicalization, close link to lexical semantics)

Data in morphology

- Unlike syntacticians, morphologists have traditionally recognized importance of *extensional* linguistic data

Data in morphology

- Unlike syntacticians, morphologists have traditionally recognized importance of *extensional* linguistic data
- In word formation, *attestedness* matters, cf. notion of *possible* vs. *existing* word, issues of lexical storage

Data in morphology

- Unlike syntacticians, morphologists have traditionally recognized importance of *extensional* linguistic data
- In word formation, *attestedness* matters, cf. notion of *possible* vs. *existing* word, issues of lexical storage
- (In syntax – except in recent "constructional" approaches – it makes no sense to distinguish between *possible* and *existing* well-formed sentences)

Data in morphology

- Unlike syntacticians, morphologists have traditionally recognized importance of *extensional* linguistic data
- In word formation, *attestedness* matters, cf. notion of *possible* vs. *existing* word, issues of lexical storage
- (In syntax – except in recent “constructional” approaches – it makes no sense to distinguish between *possible* and *existing* well-formed sentences)
- Traditionally, data in morphology come from dictionaries

Problems with dictionaries

- Underestimation of very productive, “unintentional” word formation processes

Problems with dictionaries

- Underestimation of very productive, “unintentional” word formation processes
- Overestimation of “fancy” word formation (e.g., latinate/neoclassic wf in specialized lexicon)

Problems with dictionaries

- Underestimation of very productive, “unintentional” word formation processes
- Overestimation of “fancy” word formation (e.g., latinate/neoclassic wf in specialized lexicon)
- History and contemporary language mixed

Problems with dictionaries

- Underestimation of very productive, “unintentional” word formation processes
- Overestimation of “fancy” word formation (e.g., latinate/neoclassic wf in specialized lexicon)
- History and contemporary language mixed
- Criteria for selection of entries not clear

Problems with dictionaries

- Underestimation of very productive, “unintentional” word formation processes
- Overestimation of “fancy” word formation (e.g., latinate/neoclassic wf in specialized lexicon)
- History and contemporary language mixed
- Criteria for selection of entries not clear
- No frequency information

Problems with dictionaries

- Underestimation of very productive, “unintentional” word formation processes
- Overestimation of “fancy” word formation (e.g., latinate/neoclassic wf in specialized lexicon)
- History and contemporary language mixed
- Criteria for selection of entries not clear
- No frequency information
- Very little contextual information

Problems with dictionaries

- Underestimation of very productive, “unintentional” word formation processes
- Overestimation of “fancy” word formation (e.g., latinate/neoclassic wf in specialized lexicon)
- History and contemporary language mixed
- Criteria for selection of entries not clear
- No frequency information
- Very little contextual information
- More and more dictionaries are corpus-based in any case

The importance of the past tense debate

- The English past tense debate between connectionists and defenders of the symbolic approach. . .

The importance of the past tense debate

- The English past tense debate between connectionists and defenders of the symbolic approach…
- not quite corpus-based

The importance of the past tense debate

- The English past tense debate between connectionists and defenders of the symbolic approach…
- not quite corpus-based
- and for some participants focus on morphology feels “incidental”

The importance of the past tense debate

- The English past tense debate between connectionists and defenders of the symbolic approach. . .
- not quite corpus-based
- and for some participants focus on morphology feels “incidental”
- but stressed importance of frequency data

The importance of the past tense debate

- The English past tense debate between connectionists and defenders of the symbolic approach…
- not quite corpus-based
- and for some participants focus on morphology feels “incidental”
- but stressed importance of frequency data
- and relevance of computational simulations of learning to theoretical debate

The importance of the past tense debate

- The English past tense debate between connectionists and defenders of the symbolic approach…
- not quite corpus-based
- and for some participants focus on morphology feels “incidental”
- but stressed importance of frequency data
- and relevance of computational simulations of learning to theoretical debate
- (See Albright and Hayes 2003 for a take on English past tense from a linguists’ point of view)

Corpus-based simulations of morphological learning

- Lots of recent NLP work; on the linguistic side, Goldsmith's Linguistica project, my Ph.D. work, Vito Pirrelli's SOMs (focus on inflectional paradigms, e.g., Pirrelli et al. 2003)

Corpus-based simulations of morphological learning

- Lots of recent NLP work; on the linguistic side, Goldsmith's Linguistica project, my Ph.D. work, Vito Pirrelli's SOMs (focus on inflectional paradigms, e.g., Pirrelli et al. 2003)
- Emphasis on *unsupervised* models: ultimate frontier of learning simulations

Corpus-based simulations of morphological learning

- Lots of recent NLP work; on the linguistic side, Goldsmith's Linguistica project, my Ph.D. work, Vito Pirrelli's SOMs (focus on inflectional paradigms, e.g., Pirrelli et al. 2003)
- Emphasis on *unsupervised* models: ultimate frontier of learning simulations
- Early models word-frequency-list-based, but increasing role played by context

Corpus-based simulations of morphological learning

- Lots of recent NLP work; on the linguistic side, Goldsmith's Linguistica project, my Ph.D. work, Vito Pirrelli's SOMs (focus on inflectional paradigms, e.g., Pirrelli et al. 2003)
- Emphasis on *unsupervised* models: ultimate frontier of learning simulations
- Early models word-frequency-list-based, but increasing role played by context
- Not much contact with corpus linguistics

Corpora in productivity studies

- Focus of this seminar

Corpora in productivity studies

- Focus of this seminar
- Work by Baayen and colleagues

Corpora in productivity studies

- Focus of this seminar
- Work by Baayen and colleagues
- Productivity: the “readiness” with which a wf process can form new words in a language (*-ness* vs. *-ity*, *re-* vs. *en-*)

Corpora in productivity studies

- Focus of this seminar
- Work by Baayen and colleagues
- Productivity: the “readiness” with which a wf process can form new words in a language (*-ness* vs. *-ity*, *re-* vs. *en-*)
- Early (earliest?) tradition of usage of corpora in work published in “mainstream” theoretical linguistics journals (from late eighties)

Corpora in productivity studies

- Focus of this seminar
- Work by Baayen and colleagues
- Productivity: the “readiness” with which a wf process can form new words in a language (*-ness* vs. *-ity*, *re-* vs. *en-*)
- Early (earliest?) tradition of usage of corpora in work published in “mainstream” theoretical linguistics journals (from late eighties)
- Corpus seen as word frequency list

Corpora in productivity studies

- Focus of this seminar
- Work by Baayen and colleagues
- Productivity: the “readiness” with which a wf process can form new words in a language (*-ness* vs. *-ity*, *re-* vs. *en-*)
- Early (earliest?) tradition of usage of corpora in work published in “mainstream” theoretical linguistics journals (from late eighties)
- Corpus seen as word frequency list
- Links to old tradition of lexical statistics, stylometry, authorship attribution (Baayen 2001)

Corpora in productivity studies

- Focus of this seminar
- Work by Baayen and colleagues
- Productivity: the “readiness” with which a wf process can form new words in a language (*-ness* vs. *-ity*, *re-* vs. *en-*)
- Early (earliest?) tradition of usage of corpora in work published in “mainstream” theoretical linguistics journals (from late eighties)
- Corpus seen as word frequency list
- Links to old tradition of lexical statistics, stylometry, authorship attribution (Baayen 2001)
- Less affected by later developments in corpus linguistics and corpus-based NLP

Word-formation, lexical semantics, corpora

- Recent burst of interest in semantic aspects of morphology (Lieber, 2004)

Word-formation, lexical semantics, corpora

- Recent burst of interest in semantic aspects of morphology (Lieber, 2004)
- A good moment to explore how corpora and corpus-linguistic methodology (collocational analysis, contextual approaches to meaning, emphasis on lexico-grammar) can help morphological research

The “importance of low frequency events” dilemma

- Students of word formation, by definition, trade in low frequency words

The “importance of low frequency events” dilemma

- Students of word formation, by definition, trade in low frequency words
- Very large corpora are needed to find enough rare events (e.g., in project with Lüdeling, Evert, we are studying compounding with metaphorical *obsession* – we find only 23 relevant tokens in 1.65B words German corpus)

The “importance of low frequency events” dilemma

- Students of word formation, by definition, trade in low frequency words
- Very large corpora are needed to find enough rare events (e.g., in project with Lüdeling, Evert, we are studying compounding with metaphorical *obsession* – we find only 23 relevant tokens in 1.65B words German corpus)
- Very large corpora require automated processing, and acceptance of a high degree of noise

The “importance of low frequency events” dilemma

- Students of word formation, by definition, trade in low frequency words
- Very large corpora are needed to find enough rare events (e.g., in project with Lüdeling, Evert, we are studying compounding with metaphorical *obsession* – we find only 23 relevant tokens in 1.65B words German corpus)
- Very large corpora require automated processing, and acceptance of a high degree of noise
- Automated processing is more likely to fail on low frequency events, and especially new formations!