The lexico-grammar of stance:
an exploratory analysis of scientific texts

Stefania Degaetano & Elke Teich
Universität des Saarlandes
Saarbrücken

Linguistische Profile interdisziplinärer Register (2006-2009)
Register im Kontakt (2011-2014)
Overview

• Background & Motivation
• Corpus
• Methodology
• Analysis & Results
• Discussion & Future Work
Background & Motivation

Growing interest in meaning-oriented analysis of texts

• Descriptive linguistics/corpus linguistics

• Computational linguistics
Background & Motivation

Meaning potential is associated to functions (metafunctions)

• ideational
  – expression of experience, including processes within and beyond the self and phenomena of the external world and of consciousness

• interpersonal
  – personal participation; it expresses the speaker’s role in the speech situation, his personal commitment and his interaction with others

• textual
  – concerned with the creation of text; it expresses the structure of information, and the relation of each part of the discourse to the whole and to the setting

(Halliday 1973: 351)
Background & Motivation

Understanding of interpersonal meaning is still fragmentary because

it is realized in a variety of forms

• phrasal/clausal, e.g., *it is important that, it is obligatory to*
• lexical, e.g., modal verbs, modal adverbs
• many ambiguous lexemes (connotations!)

it is extremely context-dependent (register)

• phrasal/clausal, e.g., *You should write an outline.* (British National Corpus) vs. *It is obligatory to write an outline.*
• many ambiguous lexemes (connotations!)
Background & Motivation

• Contribute to a better understanding of how interpersonal meaning is expressed

• Registers of scientific writing
  commonalities/differences across different scientific disciplines in expressing interpersonal meaning

“The notion of register is typically described as functional variation” (Quirk et al. 1985:15), i.e., variation according to type of situational context.
Corpus

Darmstadt Scientific Text Corpus (DaSciTex)

- full English journal articles (early 2000’s)
- approx. 17 million words
- tokenized, pos-tagged, lemmatized
- currently being diachronically extended (1960/70’s)

(Teich & Holtz 2009, Teich & Fankhauser 2010)
Methodology

Stance

• one kind of interpersonal meaning
• refers to how writers convey personal feelings and assessments in addition to propositional content
• three kinds of meaning associated with stance
  – epistemic (e.g. probably, it is possible to)
  – attitudinal (e.g. surprisingly, it is important to)
  – style (e.g. honestly, briefly)

(this kind of interpersonal meaning is also known under other labels: ‘evaluation’ (Hunston & Thompson 2003), ‘appraisal’ (Martin 2003), ‘hedging’ (Hyland 1996))
Methodology

Stance realized by lexico-grammatical patterns

“[...] if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it [...]” (Hunston & Francis 2000: 37)
Methodology

Examples

**it is ADJ to-INF**

*It is, however, possible to call this result into question.* (C1-Linguistics)

**it is ADJ that**

*It is clear that in some cases nesting is correlated with [...].* (C2-Biology)

**this v-link ADJ for/to/if**

*This is difficult to do for the algorithm.* (A-ComputerScience)

**make it ADJ**

* [...] two facts make it possible to classify the genes.* (C2-Biology)

**dt most ADJ n**

* [...] since they have the most important optimization potential [...].* (B3-CAD)

**evaluative noun of**

*Its main drawback lies in the difficulty of obtaining a large set [...] (B1-CompLing)*

(cf. Degaetano 2010)
Methodology

Extraction of pattern instances

Corpus Query Processor (Evert 2005)

• searches by means of regular expressions

• one very common pattern is the *it is ADJ to-INF pattern, e.g. it is easy to*

"it|It" [pos="VB.*"][][]{0,3}[pos="J.*"] "to";
Methodology

Examples

means of this technique, draw qualitative and, intended. Related works use some different technique to capture of a packet, so take advantage of multiplexing. 4.1. Model features consider that our network of queues is infinite, eliminate packets from the well as loss modelling, consider losses in the available one. Therefore, limit the number of packets of the system. First, calculate the probability multiprocessor platform, change execution schedulic conditions. Besides, simulate test environment or chaotic properties, often easiest to pick them at random. Ex
# Methodology

## Stance & meaning groups

<table>
<thead>
<tr>
<th>EPISTEMIC stance</th>
<th>ATTITUDINAL stance</th>
<th>COMPLEXITY</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POSSIBILITY</strong></td>
<td><strong>IMPORTANCE</strong></td>
<td><strong>DIFFICULT, HARD</strong></td>
<td><strong>INTERESTING, INTRIGUING</strong></td>
</tr>
<tr>
<td>possible, feasible</td>
<td>important, necessary, relevant, vital, essential, significant</td>
<td>easy, simple</td>
<td>worthwhile, worth</td>
</tr>
<tr>
<td>impossible, unfeasible</td>
<td>trivial, unimportant, unnecessary</td>
<td></td>
<td>natural, common, customary</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>reasonable, plausible</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>useful, instructive, advantageous, helpful</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sufficient, enough</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>desirable</td>
</tr>
</tbody>
</table>

(classified according to WordNet)
Analysis

Stance as expressed by the *it is ADJ to-INF* pattern

- differences / commonalities across different registers of DaSciTex in terms of stance

- Do the ‘mixed disciplines’ show differences in comparison to computer science and their ‘pure disciplines’?
## Analysis 1 – Results

### Epistemic vs. attitudinal stance

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Epistemic</th>
<th></th>
<th></th>
<th>Attitudinal</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>%</td>
<td></td>
<td>F</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>186</td>
<td>32.75</td>
<td>382</td>
<td>67.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>72</td>
<td>29.51</td>
<td>172</td>
<td>70.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>144</td>
<td>33.64</td>
<td>284</td>
<td>66.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>133</td>
<td>28.60</td>
<td>332</td>
<td>71.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td>164</td>
<td>38.86</td>
<td>258</td>
<td>61.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>129</td>
<td>32.74</td>
<td>265</td>
<td>67.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>75</td>
<td>35.38</td>
<td>137</td>
<td>64.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>153</td>
<td>36.17</td>
<td>270</td>
<td>63.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>205</td>
<td>35.59</td>
<td>371</td>
<td>64.41</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|---------------------|-------------------|-------------------------------|-------------------|--------------------------|---------------------|
Analysis 1 – Results

Summary

Within the *it is* ADJ *to*-INF pattern

- more attitudinal stance expressed by the IMPORTANCE and COMPLEXITY-group

- less epistemic stance expressed by the POSSIBILITY-group
Analysis 2 – Results

Meaning groups

![Bar chart showing the distribution of categories for different groups.](chart.png)
Analysis 2 – Results

IMPORTANCE-group

Mixed disciplines

Pure disciplines

<table>
<thead>
<tr>
<th>Course</th>
<th>A-CompSci</th>
<th>B1-Compling</th>
<th>B2-BiolInf</th>
<th>B3-CAD</th>
<th>B4-MicroElec</th>
<th>C1-Ling</th>
<th>C2-Bio</th>
<th>C3-MechEng</th>
<th>C4-ElecEng</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>12.5</td>
<td>28.28</td>
<td>28.27</td>
<td>40</td>
<td>35.55</td>
<td>27.66</td>
<td>28.3</td>
<td>36.41</td>
<td>25.87</td>
</tr>
</tbody>
</table>

23.02.2011
Analysis 2 – Results

COMPLEXITY-group

Mixed disciplines

Pure disciplines

A-CompSci | 35.39
B1-CompLing | 31.15
B2-BiolInf | 24.07
B3-CAD | 22.80
B4-MicroElec | 18.72
C1-Ling | 22.59
C2-Bio | 25.00
C3-MechEng | 18.20
C4-ElecEng | 25.17

23.02.2011
## Analysis 2 – Results

### Significant differences in DaSciTex

<table>
<thead>
<tr>
<th>corpora</th>
<th>p-value</th>
<th>signif.</th>
<th>direction</th>
<th>POSSIBILITY</th>
<th>IMPORTANCE</th>
<th>COMPLEXITY</th>
<th>OTHERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1 – A</td>
<td>3.099e-07</td>
<td>s</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B2 – A</td>
<td>5.979e-10</td>
<td>s</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B3 – A</td>
<td>&lt; 2.2e-16</td>
<td>s</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B4 – A</td>
<td>&lt; 2.2e-16</td>
<td>s</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B1 – C1</td>
<td>0.0385</td>
<td>s</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B2 – C2</td>
<td>0.8106</td>
<td>ns</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B3 – C3</td>
<td>0.07039</td>
<td>ns</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B4 – C4</td>
<td>5.099e-05</td>
<td>s</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Mixed disciplines
- B1: computational linguistics
- B2: bioinformatics
- B3: computer aided design
- B4: microelectronics

Pure disciplines
- C1: linguistics
- C2: biology
- C3: mechanical engineering
- C4: electrical engineering

A: computer science

23.02.2011
Analysis 2 – Results

Summary

Mixed disciplines

1. make more use of the IMPORTANCE-group than computer science (A)
2. bioinformatics (B2) and computer aided design (B3) similar to their pure disciplines
3. very pronounced distinctness of microelectronics (B4) (differs in the same way from A and C4)
4. less pronounced difference of computational linguistics (B1)
Analysis 3
Thing evaluated

Examples
1. *It is important to evaluate the winglets [...] (C3)*

2. *Thus, it is important to model the functionality (C4)*

3. *It is important to note that the shape [...] (C3)*

4. *At this point, however, it is important to highlight the following [...] (C4)*
## Analysis 3 – Results

\textit{important} + cognitive verb

<table>
<thead>
<tr>
<th>lexical verb</th>
<th>F</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>note</td>
<td>152</td>
<td>58.91</td>
</tr>
<tr>
<td>understand</td>
<td>18</td>
<td>6.98</td>
</tr>
<tr>
<td>consider</td>
<td>17</td>
<td>6.59</td>
</tr>
<tr>
<td>observe</td>
<td>14</td>
<td>5.43</td>
</tr>
<tr>
<td>realize</td>
<td>10</td>
<td>3.88</td>
</tr>
<tr>
<td>notice</td>
<td>9</td>
<td>3.49</td>
</tr>
<tr>
<td>recognize</td>
<td>6</td>
<td>2.33</td>
</tr>
<tr>
<td>remark</td>
<td>5</td>
<td>1.94</td>
</tr>
<tr>
<td>remember</td>
<td>5</td>
<td>1.94</td>
</tr>
</tbody>
</table>

\[ \rightarrow \textit{important} + \textit{note} \text{ different functional status} \]

\[ \rightarrow \text{formulaic expression with stylistic meaning?!} \]
## Analysis 3 – Results

**ADJ + note within the it is ADJ to-INF pattern**

<table>
<thead>
<tr>
<th>ADJ</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>important</td>
<td>152</td>
<td>51.70</td>
</tr>
<tr>
<td>interesting</td>
<td>109</td>
<td>37.07</td>
</tr>
<tr>
<td>worthwhile</td>
<td>10</td>
<td>3.40</td>
</tr>
<tr>
<td>worth</td>
<td>6</td>
<td>2.04</td>
</tr>
<tr>
<td>worthy</td>
<td>3</td>
<td>1.02</td>
</tr>
<tr>
<td>instructive</td>
<td>2</td>
<td>0.68</td>
</tr>
<tr>
<td>easy</td>
<td>2</td>
<td>0.68</td>
</tr>
<tr>
<td>significant</td>
<td>2</td>
<td>0.68</td>
</tr>
<tr>
<td>pertinent, surprising, critical, essential, useful, possible, crucial, sufficient (each occurring once)</td>
<td>8</td>
<td>2.72</td>
</tr>
</tbody>
</table>

23.02.2011
Analysis 3 – Results

Occurrences of *note* in DaSciTex

<table>
<thead>
<tr>
<th>Type of occurrence</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>it is ADJ to note</em></td>
<td>294</td>
<td>64.33</td>
</tr>
<tr>
<td><em>note</em> not within the pattern</td>
<td>163</td>
<td>35.67</td>
</tr>
<tr>
<td><em>note</em> (base form) total in DaSciTex</td>
<td>457</td>
<td></td>
</tr>
</tbody>
</table>
Analysis 3 – Results

it is ADJ to note within DaSciTex

![Bar chart showing mixed and pure disciplines]
Analysis 3 – Results

Summary

*it is ADJ to note*

• most often used with *important* and *interesting*
• basic form of *note* in DaSciTex more often used within the *it is ADJ to-INF* pattern
  ➔ seems to be used as a formulaic expression

• relatively frequently used by microelectronics (B4)
Discussion & Future Work

• investigate additional verbs occurring within the *it is* ADJ *to*-INF pattern (process types: mental, material, verbal, relational)

• investigate additional patterns to find more evidence of the tendencies of cross-disciplinary variation

• explore the constraints between *evaluative category* and *thing evaluated* for
  – potentially discriminatory effects between scientific disciplines
  – automatic attribution of the value of the *evaluative category* to the *thing evaluated*

• explore automated approaches for annotation of interpersonal expressions and probabilistic methods for corpus comparison
Discussion & Future Work

• knowledge of how evaluative patterns are constructed brings a better understanding of interpersonal meaning, e.g. stance

• the pattern approach allows a fairly easy identification of particular stance expressions in large corpora

• this knowledge may be used to improve existing systems in sentiment analysis
  – i.e., the classification approach and its extraction pattern learning algorithms (Wiebe and Riloff (2003)) and
  – the evaluative category and the thing evaluated could be automatically identified

• interpersonal meaning is context-dependent (register)
Thank you for your attention!

Any questions?
References


References


References


