Using corpus analysis to compare the explanatory power of linguistic theories: A case study of the modal load in if-conditionals

Costas Gabrielatos
Edge Hill University
costas.gabrielatos@edgehill.ac.uk

1 Motivation, background and aims

Corpus-based examinations of the extent of modal marking (modal load) in if-conditionals in the written BNC (Gabrielatos 2007, 2010: 189-229, 266-295) have revealed that they have a significantly higher modal load (ML) than average, as well as a higher ML than indirect interrogatives with if and whether, concessive conditionals with even if and whether, and non-conditional bi-partite constructions with when and whenever. Crucially, if-conditionals show a higher ML than conditionals with other subordinators (e.g. assuming). Also, the ML of if-conditionals protases (i.e. their subordinate parts) is comparable to that of the baseline, despite their being already modally marked by if.

Explanations for the emerging ML patterns are sought in the tenets of two recent linguistic theories: Construction Grammar (CxG) (e.g. Fillmore 1998) and Lexical Grammar (LG) (e.g. Sinclair 1996). The juxtaposition was motivated by the significant overlap in their tenets. Both theories take into account meaning (semantic and pragmatic) as well as lexical and grammatical factors. Constructions are symbolic units with particular features pertaining to their form and meaning. The former specify morphological, phonological and syntactic properties; the latter specify semantic and pragmatic attributes (e.g. Croft and Cruse 2004: 257-258). In LG, the unit is the lexical item, comprising the core (a word/phrase) and its collocates, semantic preference, semantic prosody and collocations (Sinclair 1996). The core difference between the two theories is that LG gives clear prominence to lexis over grammar (the component of coligation is optional), whereas CxG posits no distinction between them.

This study examines whether the ML of if-conditionals can be explained by the semantic preference of the word if (LG), or by the semantic component of conditional constructions (CxG).

2 Data and methodology

The study uses eleven random samples of 1,000 s-units from the written BNC:

- All types of constructions, providing an indication of the average frequency of modal marking in written British English – which was used as the baseline.
- Non-conditional constructions, taken collectively.
- Conditional constructions with assuming, if, in case, on condition, provided, supposing, and unless.
- Conditional-concessive constructions with even if and whether.
- Indirect interrogative constructions with if and whether.
- Constructions with when and whenever, as they are seen as combining “condition with time” (Quirk et al. 1985: 1089), and are presented as synonymous with unmodalised if-conditionals (e.g. Palmer 1990: 174-175).

The methodology combined close analysis of the samples (in order to divide the constructions into clauses), manual annotation (for modal marking), and quantitative analysis. The ML was established through the interaction of two complementary metrics: modal density and modalisation spread (Gabrielatos 2010: 50-52). Modal density (MD) is the average number of modal markings per clause, and is expressed as the number of modal markings per 100 clauses. Modalisation spread (MS) is the proportion of constructions that carry at least one modal marking, and is expressed as the percentage of modalised constructions. MD helps comparisons between samples by normalising for the complexity of the constructions in each, while MS corrects for heavily modalised constructions in the sample (see Ball 1994: 297-300). The ML of constructions is represented graphically in a scatterplot as the interaction of MD and MS values (see Figure 1). The size of similarities/differences of the MD and MS values of the constructions in focus was also examined using hierarchical cluster analysis (Gabrielatos 2010: 52-54).

The empirical comparison of the explanatory power of the two theories was motivated by the following overarching hypothesis: If different types of constructions sharing the same subordinator (particularly if) show similar ML, then this could be seen as the result of the subordinator’s semantic preference (SP), and would indicate support for LG. If constructions within the same family (particularly

---

1 This was calculated using a random sample from the written BNC, the ML of which was used as the baseline.
2 If was not included in the modal load.

---

3 An s-unit is a stretch of text delimited on either side by a sentence-boundary marker (e.g. full-stop, question mark) (Sperber-McQueen and Burnard 2007).
conditionals) show similar ML, irrespective of their subordinators, then this could be seen as the result of the construction’s semantic component, and would indicate support for CxG.

More precisely, the baseline and non-conditional constructions provided initial reference points against which the ML of the constructions in focus could be compared. Comparisons between conditional constructions helped investigate a) whether all conditional constructions have comparable ML, and, when this was not the case, b) the extent to which the ML of a conditional construction could be attributed to its subordinator. Comparisons with even-if concessive-conditionals and indirect interrogatives with if helped investigate the extent to which ML was due to the nature of the conditional construction or the word if. Comparisons with conditional-concessives and indirect interrogatives with whether provided a reference point for conditional-concessive and indirect interrogative constructions respectively — also providing further opportunities to examine the influence of subordinators on ML. The comparisons were carried out not only between whole constructions, but also between their subordinate parts. The latter comparison was deemed necessary, because the ML of subordinate parts can better reflect the SP of subordinators within the usual collocation span of 5 words.

The comparisons of MD and MS take into account the statistical significance of differences, using the log-likelihood statistic, with p≤0.05 as the threshold for statistical significance, and p≤0.01 indicating high statistical significance. The MD and MS values were also submitted to cluster analysis to reveal their progressive patterning.

3 Main results

The comparison of the ML of whole constructions (Figure 1)\(^4\) and the cluster analysis (Figure 2) revealed patterns which seem to support an explanation of ML in terms of CxG:

- If-conditionals (if\_end) have a much higher MD and MS than indirect interrogatives with if (if\_q) (p≤0.01) — and the two constructions are in completely different clusters.
- The conditional-concessives with even if (even-if\_cc) are in the same pre-final cluster with if\_end, whereas if\_q is found in the other pre-final cluster.
- The two indirect interrogative constructions (if\_q and whether\_q) cluster together, despite having different subordinators.

However,

- although most conditionals cluster together, two of them (in case, on condition) are in a different pre-final cluster;
- the two conditional-concessives (even if and whether) are in different pre-final clusters.

The examination of the ML of subordinate parts (Figure 3) and the clustering of their ML values (Figure 4), however, seem to indicate that the ML may be due to the SP of if — particularly if we consider that even if can be expected to have a different SP from if (e.g. Quirk et al. 1985: 1002).

- if\_end and if\_q have comparable ML, and cluster together.
- even-if\_cc have much lower ML than if\_end and if\_q, and are in a different major cluster.

Still, another pattern seems to provide support for CxG:

- whether\_q have comparable ML with both if\_q and share the same major cluster.

4 Brief discussion

At first glance, neither the SP of the subordinator, nor the type of construction, on their own, seem able to fully explain the results. What seems to best explain the ML patterns is their combined effect. In this light, CxG clearly demonstrates a stronger explanatory power, as a construction specifies morphosyntactic, lexical and semantic attributes (among others), which it also treats as having equal importance. This is also supported by a closer examination of issues pertaining to the ML patterns within the immediate co-text of the subordinator and the nature of LG.

The immediate co-text (subordinate parts) had to be defined grammatically, not lexically. This was because if is not a ‘free agent’. On its own, if is found in two constructions (conditionals, indirect interrogatives), and in further two as part of a multi-word subordinator: conditional-concessives (even if) and comparison clauses (as if) (Quirk et al. 1985: 1110). Therefore, a collocational analysis of if would not reveal useful patterns, as it would provide a homogenised picture of its SP in the four constructions taken together — with prominence on its collocates in if-conditionals, as they account for an estimated 85% of its occurrences (Gabrielatos 2010: 194). Finally, a collocational analysis of if aiming at establishing its semantic preference would not be posited within LG, as it does not treat function words as cores of lexical items.

The combined influence of subordinator and construction type is fully consistent with the tenets of CxG. More so, CxG accounts for the interaction between all components of a construction through the “principle of no synonymy” (Goldberg, 1995: 67): morphosyntactic (including lexical) differences between constructions lead to semantic/pragmatic differences, and vice versa.

\(^4\) In Figures 1 and 3, dotted lines indicate the MD and MS of the baseline.
5 Conclusions

The analysis has provided strong indications that CxG rather than LG can account for the ML patterns examined here. It was also shown that ML patterns are sensitive to different combinations of constructional attributes, as it would be predicted by the principle of no synonymy. This suggests that subordinators, rather than being the core of a lexical item, are better seen as one of many components defining a construction. Consequently, if a semantic attraction of the subordinator can be posited, this has to be understood as being influenced by the type of construction that the subordinator is used in. In this light, semantic preference could be more usefully treated as part of a construction’s semantic component.

References


Using corpus analysis to compare the explanatory power of linguistic theories
A case study of the modal load in *if*-conditionals

Costas Gabrielatos
Edge Hill University
Motivation

Corpus based studies of the modal load (i.e. extent of modal marking) in if-conditionals in the written BNC (Gabrielatos 2007, 2010) have revealed that they have a significantly higher modal load than ...

• average
• concessive conditionals with even if and whether,
• indirect interrogatives with if and whether,
• comparative constructions with as if and as though
• conditionals with other subordinators (assuming, in case, on condition, provided, supposing, unless).
• constructions with when and whenever (used as conjunctions)
• non-conditional constructions taken collectively
Research Question

Are the different modal load (ML) patterns due to ...

• the semantic preference of the lexical item *if*?
  ⇒ Lexical Grammar (LG)

• the semantic make-up of *if*-conditionals?
  ⇒ Construction Grammar (CxG)
Why the particular theories?

• Both take into account ...
  … meaning (semantic and pragmatic)
  … lexical and grammatical elements

• Main difference ...
  … LG gives clear prominence to lexis over grammar
  … CxG accounts for both in a balanced way
    -- in fact, it posits no distinction.
Modal Load
The interaction of two complementary metrics
Modal Density
Modalisation Spread
**Modal Density**

<table>
<thead>
<tr>
<th>Definition</th>
<th>Average number of modal markings per clause.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>Number of modal markings per 100 clauses. (%)</td>
</tr>
<tr>
<td>Utility</td>
<td>Helps comparisons between samples by normalising for the complexity of the constructions in each.</td>
</tr>
</tbody>
</table>

(Gabrielatos, 2008, 2010)

**Lexical Density:**

- The average number of content words per clause (Halliday, 2004: 654-655).
- The percentage of the tokens in a text that are content words (Ure, 1971).
Modalisation Spread

<table>
<thead>
<tr>
<th>Definition</th>
<th>Proportion of constructions that carry at least one modal marking.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>Proportion (%) of modalised constructions.</td>
</tr>
<tr>
<td>Utility</td>
<td>Corrects for heavily modalised constructions in the sample.</td>
</tr>
</tbody>
</table>

(Gabrielatos, 2010)

Spread:

- The proportion of corpus speakers who use a particular language item (Gabrielatos & Torgersen, 2009; Gabrielatos et al., 2010).
Relevant quantitative findings
(written BNC - estimations)

• On average (written BrE), we can expect...
  …about three modal markings per ten clauses
    (MD=27.7).
  …about 40% of s-units to be modalised
    (MS=40.9).

• *If*-conditionals account for about 80% of all conditional construction tokens.

• About 85% of *if tokens* are subordinators of conditional constructions.
  • The rest are subordinators of concessive-conditionals, indirect interrogatives and comparison clauses.

Written BrE is fairly heavily modalised to start with

They are excellent candidates for a case study

The word *if* is not a ‘free agent’
Modal Load comparisons
Indications that the ML is explained by the constructional nature – not the SP of *if*.
Most conditionals cluster together (irrespective of subordinator) ... 

... though not all

Indirect interrogatives cluster together (irrespective of subordinator)

Clauses of comparison cluster together (irrespective of subordinator)

Constructions with *if* do not cluster together

Conditional constructions with *if* do cluster together ...

... however ...
The ML of bi-partite constructions may not reflect the semantic preference of *if* within the usual short collocation span of 4-5 words.

Examination of ML in the subordinate part only
When we look at the immediate co-text of *if*, the ML of *if-cnd, if-q* and *as-if* is comparable. An indication that ML may be explained by the SP of *if* -- not constructional nature.

More importantly, this co-text has to be defined grammatically.

But this doesn’t explain why *even-if* should have ML significantly lower than *if-q* and *if-cnd* ...

... or why *whether-q* should have ML very close to *if-cnd, if-q* and *as-if*. 
The ML clustering of subordinate parts seems to support neither a lexis-only nor a grammar-only explanation.

Examination of the ML of matrix parts.
The matrix parts of constructions with *if* show significant differences in their ML. However, substantial differences are also seen between constructions of the same nature.
Subordinate/Matrix ML Ratio:
Shows how balanced the ML is between the subordinate and matrix parts of each construction.

if seems to have different SP in different if-constructions

Lexis + Grammar
Direct vs. Indirect *if*-conditionals (Quirk et al., 1985): ML of Protases and Apodoses
The ML of *if*-conditionals cannot be regarded as reflecting on the semantics of *if* alone, but the interaction of its semantics with the semantics of the constructions of which it is a component part.

ML seems to be explained by taking into account both ...
... lexical patterns (the SP of the subordinator)
... grammatical patterns

Which, if any, of the two theories can better accommodate this?
Lexical Grammar

Lexical Item / Extended Unit of Meaning
(Sinclair, 1996: 75, 90; Stubbs, 2009: 123-126)

The core (a word or phrase)
- Its collocates
- Its semantic prosody

Its semantic preference
- Its colligations

Lexis independent of grammar
Grammar emerges from lexical patterning

Obligatory
Optional
The focus of extended units of meaning / lexical items is content words (Sinclair, 1996; Stubbs, 2002).

⇒ No indication that ...

... *if* (a function word) would be deemed a suitable core element for an extended unit of meaning / lexical item.

... the notion of a conditional unit would be posited.

(Gabrielatos, 2010: 38)

• Collocational analysis of *if* would effectively reveal its SP within *if*-conditionals (about 80% of its uses).

• Collocational analyses of *if* within subordinate parts of particular constructions would contradict central tenets of LG.
Restoring Firthian Definitions

**Colligation**

“[F]requent co-selections of a content word and an associated grammatical frame” (Stubbs, 2002: 238).

“[T]he grammatical company a word keeps” (Hoey, 1997: 8; also Sinclair, 2004: 174).

“The statement of meaning at the grammatical level is in terms of word and sentence classes or of similar categories and of the interrelation of those categories in colligations. Grammatical relations should not be regarded as relations between words as such – between *watched* and *him* in ‘I watched him’ – but between a personal pronoun, first person singular nominative, the past tense” (Firth, 1968: 181)
A hybrid of semantic preference and colligation: “The mutual attraction holding between a sentence class ... and a semantic category” (Gabrielatos, 2007: 2).

If-conditionals can be seen as modal colligations

However ...

Not all conditionals have high ML.

The construct doesn’t fully account for the bi-partite structure of conditionals.

Semantic colligation is a reduced version of a construction.
"Conventionalised pairings of form and function" (Goldberg, 2006: 1)

"Symbolic units" with particular features pertaining to their form and meaning (Croft & Cruse, 2004: 257).

Formal properties:
morphological, phonological, lexical, syntactic

Meaning properties:
semantics, (potential) pragmatic uses

(Croft & Cruse, 2004: 258; Fillmore et al., 1988: 501; Fried & Östman, 2004: 18-21)
Accounting for the ML patterns

ML patterns result from the interaction of constructional elements / attributes ...

subordinator
  function of the subordinate part
  function of the matrix part
  nature of link between subordinate and matrix part

... and the function of the construction itself

The SP of a subordinator is influenced by the type of construction it is a component of.
Further Steps

Finalise examination of ML patterns arising from interaction of modality types and types of conditional constructions (and their subordinate and matrix parts).

In light of Herman Moisl’s presentation yesterday, establish whether non-linear cluster analysis is more suitable.
Thank you

Details: bit.ly/CG-Thesis