Lexical Cohesion Analysis of Political Speech
Web Appendix

Beata Beigman Klebanov, Daniel Diermeier, Eyal Beigman

1. WORDNET-BASED MEASURE OF RELATEDNESS

A concept in WordNet is represented by a **synset** – synonym set – a group of synonymous word senses. An orthographic word is thus not a building block of WordNet; it is merely a handle to a list of all its senses, each participating in a possibly different synset. Consider, for example, the following concepts:

(1) **Synset** : fabric (sense 1), cloth (sense 1), material (sense 3), textile (sense 1)
Gloss: Artifact made by weaving or felting or knitting or crocheting natural or synthetic fibers

(2) **Synset** : cord (sense 4), corduroy (sense 1)
Gloss: A cut pile fabric with vertical ribs; usually made of cotton

(3) **Synset** : artifact (sense 1), artefact (sense 1)
Gloss: A man-made object taken as a whole

WordNet organizes these concepts in the following taxonomy: 2 **is-a** 1 **is-a** 3. The downward-pointing arrow goes from the more general to the more specific concept (reverse of **is-a** relation); we label the nodes with one of the words whose first sense participates in the relevant synset:

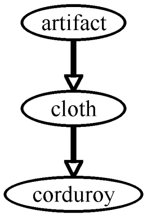

Figure 1: Corduroy-Fabric-Artifact taxonomy.

Since the relations of synonymy and hyponymy (X is-a Y) are the organizing principles of WordNet, it follows that WordNet has separate hierarchies for the different parts of speech. This is because synonymy is commonly defined via interchangeability in a suitable context, and, for grammatical reasons, different parts of speech can not be substituted for one another. In practice, WordNet contains a deep hierarchy for nouns (up to 13 levels in WordNet 2.0) and a shallow one for verbs, whereas adjectives and adverbs are only organized in synsets with glosses, without any hierarchical organization.

Measures using WordNet taxonomy are state-of-the-art in capturing semantic similarity (Jiang and Conrath, 1997; Budanitsky and Hirst, 2006). However, they would fall short of measuring cohesiveness, as, operating within a single-part-of-speech taxonomy, they cannot meaningfully compare *kill* to *death*. This is a major limitation with respect to lexical cohesion, where only about 40% of pairs marked by at least one annotator are both nouns, and less than 10% are both verbs. We thus developed a WordNet-based measure that would allow cross-part-of-speech comparisons, using glosses in addition to the taxonomy.

One family of WordNet measures are methods based on estimation of information content (henceforth, **IC**) of concepts, as proposed in Resnik (1995). He suggests that two concepts are similar to the extent that they share some content; the notion of shared content is operationalized through the lowest common subsumer of the two concepts in the taxonomy. For example, in the extract shown in figure 2, the concepts *flag*, *cloth* and *contraband* have the same pairwise similarity, which equals the information content of the concept *artifact*.

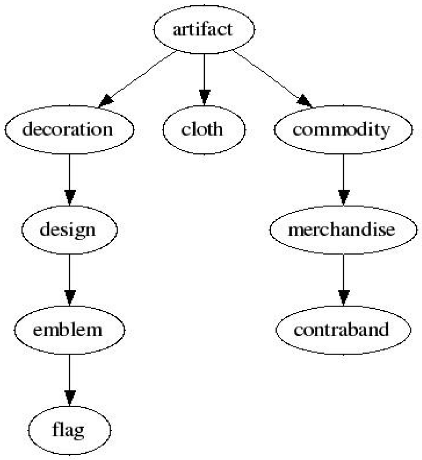

Figure 2: Flag-Cloth-Contraband excerpt from WordNet taxonomy.

Information content is usually defined using probability of occurrence, reasoning that the rarer the event the more informative it is:

$$IC(x) = -logP(x) \tag{4}$$

How does one quantify the probability of occurrence of a concept? One way would be to use synsets, and count occurrences of all word senses comprising the synset, in a sense-tagged corpus. This might work reasonably well for low and middle levels of taxonomy, under the assumption that very specific concepts (like *Union Jack* or *West Highland white terrier*) are mentioned much rarer than basic-level terms like *flag* or *dog* (Rosch, 1973). However, the top part of the taxonomycontaining very abstract concepts like *artifact* or *physical object* is typically realized by words that appear rarely in discourse. Thus, they would be assigned a high information content, rather counter-intuitively, since it seems that if two concepts merely share the property of being artifacts, they do not share much.

Resnik's key idea to overcome this problem is to count towards the concept of *artifact* every mention of something that **is-a** artifact in the taxonomy. Thus, every time the relevant sense of 'flag' is mentioned in the corpus, Resnik updates the counts for all its hypernyms as well – in this case, for *emblem*, *design*, *decoration*, *artifact*, all the way up to *entity.* This way, *artifact*, although rarely mentioned explicitly, receives high frequency and low IC value.

Resnik's method of taxonomy-based IC induction provides IC values to nominal and verbal concepts. How would one measure the informativeness of an adjective, and tell that *visible* is a property pertaining to so many things that it is uninformative, whereas *shrill* is much more narrow in application and thus more informative? Raw frequency of the relevant synsets in a corpus would not tell the difference, as both have similar low frequencies (12 and 14 in WordNet 2.0 frequency counts).

We use the observation that WordNet glosses usually list typical properties of the described concepts, which are often realized with adjectives (see *vertical* in the gloss of *corduroy* in example 2). Furthermore, those properties are mentioned at the topmost level they apply, and tend to be inherited down the taxonomy; thus, while *corduroy* has its own special characteristics, it is still *man-made*, as any artifact. Thus, properties mentioned in glosses of more general concepts are expected to be less informative, as well as properties mentioned in many different glosses.

We will count a concept's mention towards all its super-ordinates AND all words that appear in its own and its super-ordinate's glosses. This way, *visible*, which is a property of *physical object* ('a tangible and visible entity'), will get counted with each mention of something that **is-a** *physical object*, and get a low IC value, whereas *shrill* would get a high IC value since it is a property of rarely mentioned things like *whistle*, *fife*, or *stridulation*.

Now that every word in WordNet glosses is assigned an IC value, we can use glosses for comparison between word senses. Each word sense is represented as an expanded gloss – the word itself, it's own gloss, expanded, without repetition, with words appearing in the glosses of all its super-ordinate concepts, up to the top of the hierarchy. Thus, the expanded gloss of the first sense of *cloth* will contain items from glosses of artifact, unit, physical object, and entity, which is the top of the nominal hierarchy. This expanded gloss is shown in 5, with parenthesized items delimiting the contribution of the relevant glosses to the expanded gloss. If a word is repeated from a lower-level gloss, it is not added again.

(5) **Expanded Gloss of cloth#n sense 1**: cloth#n artifact#n make#v weave#v felt#v knit#v crochet#v natural#a synthetic#a fiber#n (cloth) man-made#a object#n take#v whole#n (artifact) assemblage#n part#n regard#v single#a entity#n (unit) tangible#a visible#a cast#v shadow#n (physical object) perceive#v know#v infer#v have#v own#a distinct#a existence#n live#v nonliving#a (entity)

To estimate the semantic affinity between two word senses A and B, we average the IC values of the 3 items with the highest IC in the overlap of A's and B's expanded glosses. If A^* (the word of which A is a sense) appears in the expanded gloss of B (as in the flag-cloth example before), we take the maximum between the IC(A^*) and the value returned by the 3-smoothed calculation.

To compare two words (like cloth#n and flag#n), we take the maximum value returned by pairwise comparisons of their WordNet senses. To speed the processing up, we use 5 first (most frequent) WordNet senses of each item.

2. SEMANTIC GROUPS IN THATCHER'S 1977 SPEECH

```
Group 1 (48 members):
tory thatcher labour election politics party conservative liberal britain
government manifesto socialism voter parliament socialist political
vote democratic lord callaghan british conservatism healey campaign
policy wilson exchequer opponent social elect opposition president
bevan sterling wing jenkins scargill brighton reactionary house enterprise
moderate majority win heathrow platform shirley kingdom
```

```
Group 2 (17 members):
sea boat port water sailor fishing ashore fish coast navy catch tide bait
flag labour terrify land

Group 3 (16 members):
liberal conservative conservatism wing social party socialist socialism
tory labour politics morally centre society advocate right

Group 4 (14 members):
pay money rent payment income bill buy obligation earn mortgage
rate cost store tax

Group 5 (11 members):
money economy wealth economic prosperity rich poor prosperous
inflation stagnation price

Group 6 (11 members):
truth lie confession promise true false tell say deny reality believe

The rest of the groups are shown one group per line:

month week year last ago few day thursday recent
britain country ireland nation europe state kingdom land
give take receive share get reward chance present
director executive leader manager head resign company
road way course drive narrow wheel curb
idea thought mind opinion think brain belief
speak talk tell hear listen say reply
threat danger safe risk fear dangerous threaten
reality really real ally true virtually actually
britain british lord kingdom london heathrow thatcher
family child parent newlywed home education
left wing leave right socialist instinctively
strength strong muscle strengthen courage healthy
industrial factory total production industry totaler
certainly sure yes certain indeed assure
read write writer letter dear book
stand sit standing position standard rest
fast faster easy belfast grow quickly
national nation nationalisation stagnation international nationalise
group society community member people belong
flow run go walk start move
hand carry hold touch finger firm
else nothing anyone everyone something
```

```
milk recipe food cook fruit
how answer question ask why
troops soldier force surrender regiment
want prefer wish like hope
business building build house office
fight war fighting soldier conflict
crime police victim violence accuse
great huge big vast massive
enterprising enterprise price enter prize
birmingham manchester london city glasgow
trust belief faith respect believe
fear horror panic frightening terrify
move leave go come away
union unite unionist together trade
hard effort total task try
slide decline fall rise down
crime steal bad vice conviction
deep sea platform drill water
value money price cost spend
listen hear tune sound
fall fear panic worry
important port support portray
backbone backup back background
opponent prize win match
personal property private own
life living alive live
law rule school principle
clothes look wear worn
people society woman generation
power government authority control
ride rid override overriding
undermine determined determine mine
autumn spring fall winter
flow sea stream water
economy save money spend
serviceman armed soldier force
run move movement action
mind forget think remember
reactionary act action react
create make build destroy
suppose think imagine guess
reply answer tell ask
end start begin stop
minister prime parliament thatcher
left wing "left-wing" "left-winger"
```

```
extremist extremely extreme extremism
double downgrade upgrade single
history modern century old
fund money mortgage financial
"pre-election" unelected elect election
sensible good helpful sensibly
moderate rate moderately moderation
fight win match compete
share shareholder stake hold
navy ail award force
see wait appear watch
loss win gain lose
```

REFERENCES

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 448–453.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In Moore, T. E., editor, *Cognitive Development and the acquisition of language*, pages 111–144. New York: Academic Press.