VERB POLYSEMY AND FREQUENCY EFFECTS IN THEMATIC FIT MODELING

Clayton Greenberg, Vera Demberg, Asad Sayeed (Saarland University)
{claytong,vera,asayeed}@coli.uni-saarland.de

Accurate estimates of thematic fit (e.g., is *cake* a good patient of *cut*?) can be useful both for a wide range of NLP applications and for cognitive models of human language processing difficulty, as processing difficulty is highly sensitive to semantic plausibility (McRae et al., 1998).

Research Questions. (1) Are thematically well-fitting role-fillers for more polysemous verbs (e.g., "execute killer" or "execute will") judged to be equally well-fitting as thematically well-fitting role-fillers for less polysemous verbs ("jail criminal")? (2) Is a prototypical role-filler of a polysemous verb's less-frequent sense judged to be equally well-fitting as a prototypical role-filler of the verb's more frequent sense? (3) Will a well-fitting but less frequent role-filler obtain the same rating as a more frequent but similarly-fitting role-filler?

Previous work. Existing datasets of quantitative thematic fit judgements (eg. McRae et al., 1997; Padó, 2007) do not systematically vary polysemy of verbs or frequency of role-fillers.

Stimuli. We developed a new dataset containing human judgements on 720 (role-filler, verb) pairs. We paired 48 verbs having at least two salient senses (POLYSEMOUS) each with six patient-fillers: two designed to fit well with a more frequent WordNet sense, two designed to fit well with a less frequent WordNet sense, and two designed to fit poorly with all senses. Further, we paired 48 verbs belonging to only one SynSet in WordNet (MONOSEMOUS) each with four patient fillers: two designed to fit well and two designed to fit poorly. Each pair of patient-fillers in the same polysemy-fit condition had a high frequency variant and a low frequency variant.

Filler type	**Freq.**	*whip*	*punish*
`Sense1`	high	horse	criminal
	low	stallion	outlaw
`Sense2`	high	cream	-
	low	frosting	-
`Bad`	high	party	baby
	low	gathering	fetus

Table 1. Patient-fillers for the POLYSEMOUS verb *whip* and the MONOSEMOUS verb *punish*.

Methods. Task: Rate agreement with "NOUN is something that is VERB-ed" on a Likert scale from 1 (never) to 7 (always). This avoids conflating thematic fit with role-filler frequency. For instance, *croquet* is not frequently *played*, but is always something that is *played*, never *eaten*. 159 Amazon Mechanical Turk workers completed between 1 and 8 (average of 2) surveys of 15 stimuli (6 POLYSEMOUS, 4 MONOSEMOUS, 5 filler items), 10 ratings/stimulus, 1 condition/verb/worker. The filler items were 240 pairs from the McRae et al. (1997) dataset.

Results. We found that for POLYSEMOUS verbs, good patient-fillers are not as good and bad patient-fillers are not as bad, in comparison to MONOSEMOUS verbs. Also, good patient-fillers for the more frequent sense achieve higher ratings than those for the less frequent sense. In contrast to these medium effect sizes, patient-filler frequency had a very small effect size. All results significant at $\alpha = 0.001$. We also evaluated the computational models described in Greenberg et al. (2015, NAACL) on this dataset. These showed that creating multiple prototypes per verb-role and then estimating thematic fit by comparing test role-fillers against the prototypes achieves a higher Spearman's ρ correlation with human judgements than using a single prototype. As expected, we found that POLYSEMOUS verbs benefit more from having multiple prototypes per verb-role (3 point gain) than MONOSEMOUS verbs (1 point gain), and that clustering similar prototypes is a successful middle ground for mixed polysemy datasets.

Conclusions. This thematic fit judgement dataset is the first that systematically manipulates polysemy and frequency. The judgements reveal important dependencies with implications for automatic estimation. Also, these polysemy and frequency effects help characterize the nature, identity, and weighting of linguistic and conceptual material activated by verbs and their arguments.