Verb polysemy and frequency effects in thematic fit modeling

Clayton Greenberg, Vera Demberg, and Asad Sayeed
Computational Linguistics and Phonetics / M^{2}CI Cluster of Excellence
Saarland University
66123 Saarbrücken, Germany
{claytong,vera,asayeed}@coli.uni-saarland.de

Abstract

While several data sets for evaluating thematic fit of verb-role-filler triples exist, they do not control for verb polysemy. Thus, it is unclear how verb polysemy affects human ratings of thematic fit and how best to model that. We present a new dataset of human ratings on high vs. low-polysemy verbs matched for verb frequency, together with high vs. low-frequency and well-fitting vs. poorly-fitting patient role-fillers. Our analyses show that low-polysemy verbs produce stronger thematic fit judgements than verbs with higher polysemy. Role-filler frequency, on the other hand, had little effect on ratings. We show that these results can best be modeled in a vector space using a clustering technique to create multiple prototype vectors representing different "senses" of the verb.

1 Introduction

Being able to accurately estimate 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 human processing difficulty is highly sensitive to semantic plausibilities (Ehrlich and Rayner, 1981).

Previous studies obtained quantitative thematic fit data by asking human participants to rate how common, plausible, typical, or appropriate some test role-fillers are for given verbs on a scale from 1 (least plausible) to 7 (most plausible) (McRae et al., 1998; Binder et al., 2001; Padó, 2007; Padó et al., 2009; Vandekerckhove et al., 2009). For example, as part of the McRae et al. (1998) dataset, the thematic fit of the noun "principal" as the patient of the verb "dismiss" is 2.0 out of 7.0. As an agent, its rating is 6.3. The McRae et al. (1998) dataset has a total of 720 verb-noun pairs (146 different verbs) with typicality ratings. The Padó (2007) dataset includes 18 verbs as well as up to twelve nominal arguments, totalling 207 verb-noun pairs. The verbs and nouns were chosen based on their frequent occurrence in the Penn Treebank and FrameNet.

While these datasets are very useful, e.g. for evaluating automatic systems for estimating thematic fit via correlations with these human judgements, they do not systematically vary polysemy of verbs or frequency of role-fillers. Further, it is unclear what effect polysemy and frequency have on thematic fit judgements. We thus ask: (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) Finally, will a well-fitting but less frequent role-filler obtain the same rating as a more frequent but similarly-fitting role-filler?

The answers to these questions have implications for modeling thematic fit. An increasingly common method for determining the fit between a verb and its argument involves calculating typical role-fillers of that verb, calculating a centroid (or average) over the most typical role-fillers in a vector space model, and then calculating the similarity between the centroid

Proceedings of CMCL 2015, pages 48–57,
Denver, Colorado, June 4, 2015.

and the proposed role-filler via a similarity measure. Arguments that have high similarities with the prototypical centroid vector representing most common role-fillers for a given verb-role combination are then asserted to have good thematic fit (Baroni and Lenci, 2010; Erk, 2012).

This conceptualization, however, assumes that there is a single type of most typical filler for a role and that all good fillers will be distributionally similar. This assumption leads to problems when this process is to be applied to ambiguous verbs; when a verb has many different senses, there can exist typical role-fillers for each sense which are all highly suitable role-fillers for the given role but are distributionally very different from one another. This means that the calculated prototypical role-filler will be a mixture of the arguments that are typical role-fillers for the main senses of the verb.

Greenberg et al. (2015) addressed this problem by clustering the most common role-fillers in order to represent the prototypes of each of the verb senses. They found that better correlations with human judgements on the Padó (2007) and McRae et al. (1998) datasets are achieved when calculating the maximal cosine similarity for a candidate role-filler with respect to the prototypical role-fillers of each word sense.

Thus, their model represents the similarity to the most similar prototype role-filler, which means that a good role-filler of an infrequent verb sense could get the same level of ratings as a good role-filler of a frequent verb sense. There exists, however, currently no data to assess whether this is desirable behavior, or whether we, in fact, need a model that calculates similarity to prototypical role-fillers *and* takes into account verb sense frequencies.

2 Thematic fit modeling

Quantifications of thematic fit are a ternary relation between a verb, a semantic role, and a role-filler. For example, given human judges, we would expect *cake* to be a highly-rated patient-filler of *cut*, but we would expect *cake* to be a poorly-rated agent-filler of *cut*. There have been multiple attempts to model thematic fit judgements. The goal is generally to estimate a probability for a thematic role-filler given a verb and a role. However, due to data sparseness, it is not possible to estimate this probability directly. Existing approaches estimate a candidate noun's thematic fit via its similarity to typical role-fillers that have been observed. Similarity between the candidate noun and prototypical fillers is thereby assessed via WordNet classes (e.g., Resnik, 1996; Padó et al., 2009), or by cosine similarities in a vector space model (e.g., Baroni and Lenci, 2010; Erk, 2012). However, vector space methods achieve better coverage than WordNet class methods (Erk et al., 2010).

In vector-space modeling approaches like the one used in this paper, the calculation of a thematic fit for a verb-role-noun triple proceeds though the identification of a prototype vector of a verb's role-fillers. The prototype vector is constructed from the representations of words that are previously known to be typical role-fillers for that verb. For example, we might identify typical patient-fillers of *cut* to be *meat*, *budget*, *paper*, and so on. A geometric measure such as cosine similarity is used to compare the vector for the candidate role-filler with the prototype vector.

2.1 Distributional memory vector space models

The models evaluated in this paper (TypeDM, SDDM, and SDDMX) are based on the distributional memory (DM) framework originally promulgated by Baroni and Lenci (2010). DM is a generalized, broad-coverage, unsupervised model for representing linguistic relationships in a very high-dimensional vector space. A DM is an order 3 tensor, two of whose axes are words and one of whose axes is a syntactic or semantic link between words. In other words, a cell of a DM represents a tuple $< w_1, \text{link}, w_2 >$, and the value contained in that cell is an adjusted frequency count—here, local mutual information (LMI; $O_{FRV} \log \frac{O_{FRV}}{E_{FRV}}$, where O and E are the observed and expected frequencies of filler F, role R, and verb V appearing together). Using a structured vector space model is crucial for modeling thematic fit (as we need to distinguish between explicit roles, e.g. typical agents vs. patients of a verb).

The **TypeDM** model[1] (Baroni and Lenci, 2010) is constructed from the ukWaC, BNC, and WaCkype-

[1] Available at `http://clic.cimec.unitn.it/dm/`.

dia corpora. In TypeDM, the links represent both connections between words in the corpora found via the dependency parser MaltParser (Nivre et al., 2007) and further semantic dependencies derived from these connections via hand-crafted rules.

An alternative way of constructing DMs was proposed by Sayeed and Demberg (2014), where links between words are derived directly from SENNA, a neural network-based semantic role labeller (Collobert and Weston, 2007; Collobert et al., 2011). This DM is called SENNA-DepDM, or **SDDM** for short in this paper. Unlike TypeDM, the links in this tensor are not processed by hand-crafted rules.

SDDMX[2] is a version of SDDM with one expansion: it includes additional links between role-fillers that are found to be related via a verb. Both SDDM and SDDMX are trained on ukWaC and the BNC.

Greenberg et al. (2015) tested TypeDM, SDDM, and SDDMX on multiple datasets of human judgements for agent, patient, location, and instrument roles. They used multiple models and datasets because robustness of trends across these different configurations lends support to their generality. They found that the methods tested had comparable performance across the three models, with TypeDM outperforming considerably on the McRae et al. (1998) agent/patient dataset and SDDMX likewise on locations. We included TypeDM, SDDM, and SDDMX in our experimental evaluation on the new dataset to allow similar cross-model analysis.

2.2 Modeling verb senses

While prior vector space models for thematic fit have ignored verb polysemy, Greenberg et al. (2015) recently proposed to partition the "typical" role-fillers of a verb like "observe" such that each partition reflects typical role-fillers of separate senses of the verb.

In that work, they compared the traditional method of representing a prototypical role-filler by calculating a single $Centroid$ from a verb's 20 highest-LMI role-fillers with three other thematic fit estimation methods: $OneBest$, in which the cosine is taken separately with all of the 20 highest-LMI fillers and the best cosine is reported; $2Clusters$, in which the 20 fillers are partitioned into two clusters and the best fit is taken from the corresponding prototypes; and $kClusters$, in which the 20 fillers are dynamically partitioned into three or more clusters using NLTK's (Bird et al., 2009) group-average agglomerative clustering package and using the Variance Ratio Criterion (Caliński and Harabasz, 1974) as a stopping criterion for partitioning. They concluded that variable clustering ($kClusters$) provides gains in thematic fit modeling over the other methods, suggesting a need to take into account verb polysemy with respect to thematic roles in order to model human judgements more accurately. Also, since their clustering methods helped patients much more than agents, they successfully reproduced the previously known notion that patients are more specific to individual senses of a verb than agents.

3 Methods and stimuli

In this work, we describe a novel dataset of thematic fit judgements that systematically varies verb polysemy and role-filler frequency. Then, we evaluate the automatic thematic fit estimation methods from Greenberg et al. (2015) on this dataset. If verb polysemy and filler frequency can be shown to affect human thematic fit judgements, these results would suggest certain desirable traits for automatic systems and provide evidence for or against the claims made by Greenberg et al. (2015). In addition to whether the factors of polysemy and frequency are associated with shifts in the rating scale, we also would like to know how these shifts change at both scale extrema, whether good role-fillers of different verb senses receive relatively equal ratings, and how an automatic thematic fit estimation system with prototype clustering handles different types of verbs with respect to these manipulations.

We begin with the necessary evil of operationalizing polysemy. It is probably impossible to prove without a doubt that a certain verb has only one meaning, usage, etc. However, a binary classification between less polysemous and more polysemous verbs is certainly attainable, even if the boundary is not beyond reproach. For our purposes, we will define a verb as MONOSEMOUS if its lemma is a member of only one SynSet in WordNet (Fellbaum, 1998). Hence, a verb is POLYSEMOUS if its lemma

[2]SDDM, SDDMX, and this paper's dataset are available at `http://rollen.mmci.uni-saarland.de/`.

is a member of more than one SynSet in WordNet. A possible confound when manipulating polysemy is verb frequency, as higher frequency words in general tend to be more polysemous. We control for verb frequency by selecting POLYSEMOUS verbs to match the frequency of the MONOSEMOUS verbs in our dataset as much as possible. Furthermore, we systematically vary the frequency of the role-fillers, i.e., selecting a high and low frequency noun in each condition.

3.1 Task format and template

Since Greenberg et al. (2015) were able to confirm that patients are more specific to individual senses of a verb than agents, we decided to focus on patient role-fillers in our new dataset, thus emphasizing the effects of polysemy. For patient-fillers, both McRae et al. (1998) and Padó (2007) used questions of the form *"How common is it for a NOUN to be VERB-ed?"* to elicit judgements for their datasets. But consider the example: *"How common is it for croquet to be played?"* Since croquet is not a very common game, we would expect the rating in response to this question to be relatively low. But, intuitively, *croquet* is an excellent patient-filler for *play*. So, instead, we decided to ask participants to rate how much they agree with statements of the form *"A NOUN is something that is VERB-ed"* (template for non-human patient-fillers) and *"A NOUN is someone who is VERB-ed"* (template for human patient-fillers) on a Likert scale from 1 (never) to 7 (always). We chose this construction as our template because it does not use any technical terms and avoids conflating absolute frequency of the verb with conditional probability of the patient-filler, e.g. croquet is always something that is played, so it should receive a high rating.

3.2 Selection of experimental items

Given that MONOSEMOUS verbs are far less plentiful than POLYSEMOUS ones, we first selected the MONOSEMOUS verbs. To start, we filtered the 500,000 most frequent tokens in COCA (Davies, 2008) for parts of speech starting with `v` (verbs sorted by descending frequency). Then, using the WordNet lemmatizer as part of NLTK (Bird et al., 2009), we lemmatized the verbs, combined the duplicate entries that arose from multiple inflected forms, and then filtered out all lemmata that were part of multiple SynSets. The top 48 most frequent MONOSEMOUS verbs that were acceptable in our template constructions were compiled into a list. These vary in frequency from "thank" (82987 occurrences) down to "sample" (1275 occurrences).

Then, by querying COCA with the trigram *VERB* `[at*] [nn*]`, we obtained a list of excerpts from the corpus in which the verb was followed by a determiner (article) and then a noun. This targeted patient-fillers, since in English, they usually appear right after the verb. Therefore, the results of this query formed a list of candidate patient-fillers sorted by cooccurrence frequency. One particularly well-fitting patient-filler was selected from this list, giving priority to the higher (more frequent) entries. After this, using Roget's 21st Century Thesaurus accessed through `http://www.thesaurus.com`, we selected a very similar but less frequent version of the patient-filler. The relevant unigram patient-filler frequencies were obtained by querying a version of the 500,000 most frequent tokens in COCA that was filtered for only nouns and lemmatized using the WordNet lemmatizer as part of NLTK. The median ratio of the high frequency patient-fillers to their low frequency counterparts was 9.912.

Once the MONOSEMOUS verbs were finalized, we compiled the POLYSEMOUS verbs. First, we generated the same list from which the MONOSEMOUS verbs were selected, except that instead of filtering out lemmata that were part of more than one SynSet, we filtered out lemmata that were part of fewer than three SynSets. While this is stronger than our initial definition of POLYSEMOUS, we wanted to make sure that polysemy is effectively manipulated. Then, beginning at the frequency of each MONOSEMOUS verb, we looked for a verb as close in frequency as possible to that MONOSEMOUS verb that had at least two significantly contrasting, transitive senses according to experimenter intuition, confirmed by corresponding SynSets in WordNet, giving priority to verbs that were members of many SynSets. The median number of WordNet SynSets belonging to each of these 48 POLYSEMOUS verbs was 7. The frequencies of the POLYSEMOUS verbs varied from "started" (80898 occurrences) down to "scratched" (1465 occurrences).

The same format trigram COCA queries aided this POLYSEMOUS verb selection process as well as the selection of a high frequency good patient-filler for each of two senses. Priority was still given to those nouns with greatest cooccurrence with the verb, but the two-sense requirement made this more difficult. Low frequency versions of these good patient-fillers were analogously selected using the thesaurus. The more frequent of the two experimental senses of these POLYSEMOUS verbs, according to SynSet ordering in WordNet, was labelled as `Sense1` and the less frequent was labelled `Sense2`. The median ratio of the high frequency patient-fillers to their low frequency counterparts was 7.335 for `Sense1` and 10.288 for `Sense2`.

To investigate multiplicative as well as additive adjustments to the thematic fit rating scale, we needed to determine bad patient-fillers explicitly. For this we randomly shuffled the good role-fillers for the MONOSEMOUS verbs and paired them with each MONOSEMOUS verb again. If the thematic fit of a randomly assigned pair of bad patient-fillers was too good, possibly because the verb had coincidently been paired with its good patient-fillers again, a swap was made. To ensure that polysemy and other idiosyncrasies of the selected patient-fillers for MONOSEMOUS verbs were controlled, we used a random ordering of the patient-fillers for the MONOSEMOUS verbs also as the bad patient-fillers for the POLYSEMOUS verbs. Once again, swaps were made if the thematic fit of a randomly assigned pair of bad patient-fillers was too good. Note that another way to obtain bad role-fillers would have been to invert the animacy and/or concreteness of the good role-fillers. However, since this study is concerned with scalar thematic fit judgements as opposed to hard classifications, we thought that the variation in thematic fit arising from randomly selecting bad fillers would be more appropriate.

To summarize the experimental items, this dataset has 48 MONOSEMOUS verbs each with frequency-contrasting pairs of good and bad patient-fillers. Also it has 48 POLYSEMOUS verbs each with frequency-contrasting pairs of good patient-fillers for `Sense1`, good patient-fillers for `Sense2`, and bad patient-fillers. In Table 1, we show the selected patient-fillers for the POLYSEMOUS verb *whip* and the MONOSEMOUS verb *punish*.

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

Table 1: Example items from our thematic fit dataset.

3.3 Fillers

In order to evaluate consistency with the *"How common is it..."* format and also to identify excessively divergent responses, we adapted 240 (patient-filler, verb) pairs from McRae et al. (1998) as a counterpart to our novel experimental items. To select these pairs, we excluded all verbs that appeared as experimental items, scored each remaining pair using the sum of the COCA unigram frequencies of the verb and patient-filler, and selected the 240 highest scoring pairs. Note that because of this procedure, the verbs that were selected as fillers did not necessarily appear with all of their role-fillers from the McRae et al. (1998) dataset.

3.4 Experimental setup

In order to prepare the 480 total (patient-filler, verb) pairs for inclusion in a human experiment, we rewrote each verb by hand in its past participle form and each patient-filler by hand in its singular form with an appropriate (possibly null) determiner. Also, each patient-filler was hand tagged with a `+human` or `−human` feature. That way, each (patient-filler, verb) pair could be felicitously entered into the non-human-filler template or the human-filler template.

We obtained participants for this study using Amazon Mechanical Turk. For a survey consisting of six POLYSEMOUS items, four MONOSEMOUS items, and five filler items, counterbalanced for condition and question order, a worker was paid $0.15. Workers were restricted such that they were not allowed to rate a verb in more than one condition. So, each worker could complete a maximum of eight surveys. A total of 159 workers participated, and each sentence was rated by 10 different workers.

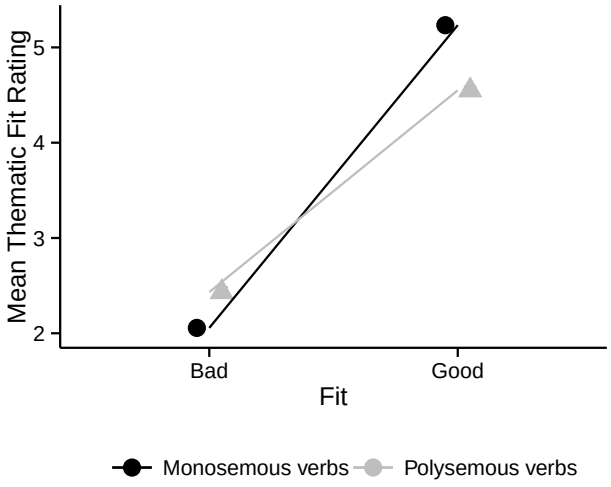

Figure 1: Interaction between Fit and $Polysemy$.

4 Results

The Spearman's ρ correlation between the human judgements we obtained on our filler items and the human judgements obtained by McRae et al. (1998) is 0.753.

Our highest level experimental analysis was a factorial ANOVA with "hc3" correction as suggested by Long and Ervin (2000), which had three, between participant, binary factors: $Polysemy$, experimenter judgement (Fit), and $Frequency$ (binned). This analysis provided two important results. First, it empirically confirmed the choices of our experimental patient-fillers, which were designed to fit either very well or poorly. This effect of Fit was significant and very large: $F(1, 4668) = 3029.692$, $p < 2 \times 10^{-16}$, $\eta_p^2 = 0.394$.

Second, there was a significant $Polysemy * Fit$ interaction, summarized visually in Figure 1, $F(1, 4668) = 125.729$, $p < 2 \times 10^{-16}$, $\eta_p^2 = 0.026$. Namely, for POLYSEMOUS verbs, bad patient-fillers were not as bad (POLYSEMOUS: $M = 2.43$, $SD = 1.56$ versus MONOSEMOUS: $M = 2.06$, $SD = 1.38$) and good patient-fillers were not as good (POLYSEMOUS: $M = 4.55$, $SD = 1.65$ versus MONOSEMOUS: $M = 5.23$, $SD = 1.44$). We used two-tailed Welch t-tests on both bad patient-fillers, $t(1813.212) = 5.4756$, $p = 4.968 \times 10^{-8}$, Cohen's $d = 0.173$, and good patient-fillers, $t(2139.706) = 11.3243$, $p < 2.2 \times 10^{-16}$, Cohen's $d = 0.272$, to confirm that these differences were significant. Finally, we found significant, but very small, main effects of $Polysemy$, $F(1, 4668) = 16.175$, $p = 5.87 \times 10^{-5}$, $\eta_p^2 = 0.003$, and also $Frequency$, $F(1, 4668) = 11.184$, $p = 0.000832$, $\eta_p^2 = 0.002$ on how people generally rated thematic fit.

Then, we ran four follow-up 2×2 factorial ANOVAs with "hc3" correction, each holding a $Polysemy$ or Fit condition constant. First, for good patient-fillers, both $Polysemy$, $F(1, 2830) = 117.761$, $p < 2 \times 10^{-16}$, $\eta_p^2 = 0.040$, and $Frequency$, $F(1, 2830) = 8.670$, $p = 0.00326$, $\eta_p^2 = 0.003$ were significant. Second, for bad patient-fillers, $Polysemy$ was significant, $F(1, 1838) = 29.997$, $p = 4.92 \times 10^{-8}$, $\eta_p^2 = 0.016$, but $Frequency$ was not, $F(1, 1838) = 2.524$, $p = 0.112$, $\eta_p^2 = 0.001$. That $Frequency$ has a significant effect on good role-fillers but not on bad ones makes intuitive sense. After all, a less frequent version of a poorly-fitting role-filler should fit poorly to approximately the same degree.

Third, for POLYSEMOUS verbs, Fit was significant, $F(1, 2803) = 1054.885$, $p < 2 \times 10^{-16}$, $\eta_p^2 = 0.273$, but $Frequency$ was not, $F(1, 2803) = 2.866$, $p = 0.0906$, $\eta_p^2 = 0.001$. Fourth, for MONOSEMOUS verbs, both Fit, $F(1, 1865) = 2373.263$, $p < 2 \times 10^{-16}$, $\eta_p^2 = 0.560$, and $Frequency$, $F(1, 1865) = 11.105$, $p = 0.000878$, $\eta_p^2 = 0.006$, were significant. That $Frequency$ has a significant effect on MONOSEMOUS verbs but not on POLYSEMOUS ones appears to be indicative of a characteristic of low polysemy. These verbs produce such a strong expectation for certain role-fillers that even role-fillers that are semantically very similar but less frequent are deemed worse-fitting. POLYSEMOUS verbs, on the other hand, are more flexible than MONOSEMOUS verbs for fitting with less frequent role-fillers.

While verb frequency was very closely controlled in our stimuli via experimental design, we also ran a linear mixed effects model with thematic fit as a response variable and POLYSEMY*FIT+LOGVERBFREQ+FREQUENCY as predictors (with random intercepts under participant and item, as well as random slopes for POLYSEMY and FIT under both participants and items). The linear mixed effects model confirmed all results from the factorial ANOVAs, and furthermore

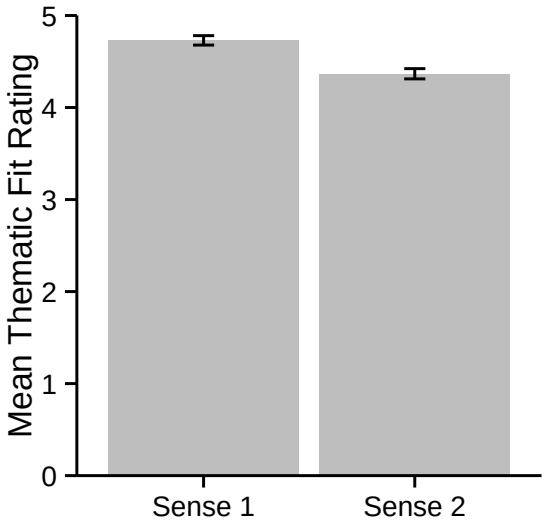

Figure 2: More frequent vs. less frequent senses of POLYSEMOUS verbs.

showed that our matching of verb frequencies in the experimental design was effective: LOGVERBFREQ did not explain away the effects of verb polysemy.

Next, we compared good patient-fillers for the two predetermined senses for the POLYSEMOUS verbs. A Factorial ANOVA with "hc3" correction and with *Sense* and *Frequency* as between participant factors indicated that there was a significant main effect of *Sense*, $F(1, 1881) = 23.076$, $p = 1.68 \times 10^{-6}$, $\eta_p^2 = 0.012$. Neither *Frequency*, $F(1, 1881) = 3.024$, $p = 0.0822$, $\eta_p^2 = 0.002$, nor the $Sense * Frequency$ interaction, $F(1, 1881) = 1.386$, $p = 0.2392$, $\eta_p^2 = 0.001$ was significant. A two-tailed Welch t-test confirmed that good patient-fillers for the more frequent sense of these POLYSEMOUS verbs were rated significantly higher ($M = 4.73$, $SD = 1.58$) than good patient-fillers for the less frequent sense ($M = 4.37$, $SD = 1.70$), $t(1868.449) = 4.7985$, $p = 1.725 \times 10^{-6}$, Cohen's $d = 0.254$, as shown in Figure 2. Therefore, while the unigram frequencies of the patient-fillers do not have an impact when comparing senses of the same verb, the frequencies of the senses themselves do have an effect.

Finally, in Table 2, we give the results of running the four automatic thematic fit scoring methods from Greenberg et al. (2015) on SDDM, SDDMX, and TypeDM and calculating the correlation with the human judgements we obtained on the experimental (role-filler, verb) pairs. For the *kClusters* method, 10 was set as the maximum number of clusters. In Table 3 we break down the TypeDM correlations by verb type. Note that the ALL column in Table 3 includes the filler items, but Table 2 does not.

	SDDM	SDDMX	TypeDM
Centroid	0.406	0.448	0.528
2Clusters	0.448	0.476	0.539
OneBest	0.509	0.531	0.544
kClusters	0.520	0.535	0.548

Table 2: Spearman's ρ values for correlation with MTurk judgements on experimental items.

	POLY.	MONO.	FILLERS	ALL
Centroid	0.405	0.655	0.313	0.464
2Clusters	0.442	0.642	0.311	0.474
OneBest	0.447	0.641	0.223	0.452
kClusters	0.432	0.669	0.304	0.479

Table 3: Spearman's ρ values for TypeDM correlation with MTurk judgements by verb type.

5 Discussion

The reasonably high correlation between our human judgements and those from McRae et al. (1998) is encouraging and provides a possible upper-bound on computational models of thematic fit as well as a human annotator agreement score for our study.

Since the *Fit* factor was experimentally designed to have an effect on ratings, it is unsurprising that there was an effect. But it is surprising that the *Polysemy* and *Frequency* effect sizes are much smaller than those of *Fit* and the interaction. This suggests that humans do not have such a varying process for assessing thematic fit for POLYSEMOUS versus MONOSEMOUS verbs. Therefore, these judgements further motivate clustering as part of an automatic thematic fit scoring system because clustering minimizes the effects of highly contrastive senses.

Overall, the interaction between *Polysemy* and *Fit* showed that in the case of POLYSEMOUS verbs, it is harder to achieve extremely low or high thematic fit. Only one sense needs to be relevant for a

role-filler to achieve a somewhat high score, but the inability to fit well with all senses may block a good role-filler from achieving the highest possible score.

For the comparison of the two experimental senses for the POLYSEMOUS verbs, it is important to note a terminological subtlety. Our *Frequency* factor, which was found not to have a significant effect, is based on the *unigram* frequency of the role-filler, while the *Sense* factor, which was found to have a significant effect, is based on the relative frequency of that sense, which could be estimated using the (skip) *bigram* frequency of the verb with the role-filler. Since these bigram frequencies affect thematic fit ratings, automatic thematic fit estimation systems that analyze the frequency distribution of senses are likely to perform better than those that do not.

Table 2 reproduces the trends in correlations observed in Greenberg et al. (2015) on our new dataset. Again, we see that the trends occur on each of the DM models, which shows their generality. But, by breaking down the dataset by verb type, we can see a clearer picture of the strengths and weakness of the different scoring methods. For instance, the *OneBest* method achieves the best performance on POLYSEMOUS verbs, but worsens performance on MONOSEMOUS verbs. We can attribute this difference to a trade-off between negative impacts of polysemy and noise. Namely, for MONOSEMOUS verbs, the negative impact of noise is greater than the negative impact of polysemy, and vice versa for POLYSEMOUS verbs. Clustering, however, achieves the greatest correlation with human judgements on mixed polysemy datasets presumably by avoiding the greater negative effect for each verb.

As an example, consider the MONOSEMOUS verb *obey*. *kClusters* put the patient-fillers of *obey* into nine clusters: [[*injunction*], [*will*], [*wish*], [*limit*], [*equation*], [*master*], [*law, rule, commandment, principle, regulation, teaching, convention*], [*voice, word*], [*order, command, instruction, call, summons*]]. Due to a large number of singleton clusters, each cluster is quite pure. Hence, the noise has been neutralized. Similar role-fillers are still smoothed together, but no strongly dissimilar ones are averaged.

In contrast, *kClusters* put the patient-fillers for the POLYSEMOUS verb *observe* into six clusters: [[*day*], [*silence*], [*difference, change*], [*object, star, bird*], [*effect, phenomenon, pattern, behaviour, practice, behavior, reaction, movement, trend*], [*rule, custom, law, condition*]]. Now, there are only two singleton clusters, and the largest cluster is quite noisy. Each of the clusters except the largest happens to correspond uniquely to a WordNet SynSet, so the polysemy has been addressed, but not the noise. However, polysemy was more important than noise for this verb. We also note that the number of clusters, usually between six and nine, is not particularly informative about polysemy and has much more to do with noise in the set.

Finally, to explain the sharp discrepancy in performance between fillers and experimental items, recall that our main experiment had three independent variables: *Polysemy*, *Frequency*, and *Fit*. Both levels of *Polysemy* enjoyed the same positive effect when moving from the *Centroid* to *kClusters*. *Frequency* had a very small effect. This just leaves *Fit*. For each of our experimental verbs, we ensured that there was a pair of good role-fillers and a pair of bad role-fillers. The McRae et al. (1998) dataset did not ensure that there was a mix of good and bad role-fillers for each verb. Additionally, our filler item selection procedure did not always include every available role-filler for a given verb. If the selected role-fillers are either all good or all bad, these points "vote", during the Spearman's ρ calculation, to minimize all distinctions (good and bad) that the model makes. The more of these verbs we have, the flatter our model becomes and the less we will be able to see. But, none of our experimental items had this problem.

6 Conclusions and Future Work

We developed a new substantial dataset of thematic fit judgements: 720 verb-noun pairs, each judged by 10 Amazon Mechanical Turk workers. Our dataset contains 48 MONOSEMOUS and 48 POLYSEMOUS verbs, matched for frequency. For each of the POLYSEMOUS verbs, it has a total of six patient-fillers: two good for `Sense1`, two good for `Sense2`, two bad, each pair with contrasting frequencies. The MONOSEMOUS verbs in our dataset have a total of four patient-fillers: two good and two bad, each pair with contrasting frequencies. This dataset constitutes the first thematic fit judgement dataset that systematically manipulates polysemy and frequency.

We found that human judgements of thematic fit are affected by the number of senses that a verb has (good role-fillers for MONOSEMOUS verbs are judged better than those for POLYSEMOUS verbs, and bad role-fillers are judged worse for MONOSEMOUS verbs than for POLYSEMOUS verbs), and that this effect cannot be explained away by the verb's frequency. This effect may reflect the different levels of constraint that a MONOSEMOUS vs. POLYSEMOUS verb exerts on its arguments. A further important finding was that the frequency of a role-filler has little influence on thematic fit judgements. This supports the notion that semantic similarity and thematic fit are extremely important notions for modeling thematic fit well.

We then evaluated distributional memory models and computational estimation methods on this dataset, comparing methods that can account for verb polysemy by clustering most typical fillers ($kClusters$) to methods that assume a single verb sense ($Centroid$). Our results show that the method that allows for representing verb polysemy consistently outperforms the traditional single-centroid method by Baroni and Lenci (2010). As expected, the most substantial improvements are achieved for POLYSEMOUS verbs, but we also found that model performance on MONOSEMOUS verbs was not hurt by using the $kClusters$ method.

The data we collected also suggests that both the probability of the verb sense and the similarity of a role-filler to a prototypical argument for a specific verb sense play a role in human thematic fit judgements: this explains why highly prototypical role fillers for MONOSEMOUS verbs get significantly higher thematic fit judgements than highly prototypical role-fillers for the most frequent verb sense of a POLYSEMOUS verb, and why, in turn, highly prototypical role-fillers for a less frequent verb sense get again significantly lower thematic fit judgements in comparison.

A model implementing this would conceptually estimate thematic fit in terms of a noun's surprisal given the verb ($-\log P(filler|verb)$), thereby using the semantic vector space as a back-off model in order to handle rare or unseen combinations of verbs and their arguments. The importance of this is highlighted by our result that noun frequency had little effect on thematic fit judgements. In all, polysemy, frequency, and thematic fit are intertwined in a complex web of dependencies, but the more carefully we obtain human judgements, the more equipped we are to build highly accurate computational models.

Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102: "Information Density and Linguistic Encoding." Also, the authors wish to thank the four anonymous reviewers whose valuable ideas contributed to this paper.

References

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Binder, K. S., Duffy, S. A., and Rayner, K. (2001). The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language*, 44(2):297–324.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics—Simulation and Computation*, 3(1):1–27.

Collobert, R. and Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. In *Annual Meeting—Association for Computational Linguistics*, volume 45, page 560.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Davies, M. (2008). The corpus of contemporary american english (COCA): 400+ million words, 1990-present. Available online at `http://www.americancorpus.org`.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K., Padó, S., and Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Fellbaum, C. (1998). *WordNet: an electronic lexical database*. Wiley Online Library.

Greenberg, C., Sayeed, A., and Demberg, V. (2015). Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics – Human Language Technologies*, Denver, USA.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Padó, U. (2007). *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. PhD thesis, Saarland University.

Padó, U., Crocker, M. W., and Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.

Sayeed, A. and Demberg, V. (2014). Combining unsupervised syntactic and semantic models of thematic fit. In *Proceedings of the first Italian Conference on Computational Linguistics (CLiC-it 2014)*.

Vandekerckhove, B., Sandra, D., and Daelemans, W. (2009). A robust and extensible exemplar-based model of thematic fit. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 826–834.