Empirical Comparisons of MASC Word Sense Annotations

Gerard de Melo[1], Collin F. Baker[1], Nancy Ide[2], Rebecca J. Passonneau[3], Christiane Fellbaum[4]

1: International Computer Science Institute, Berkeley, CA, USA – {demelo,collinb}@icsi.berkeley.edu
2: Vassar College, Poughkeepsie, NY, USA – ide@cs.vassar.edu
3: Columbia University, New York, NY, USA – becky@cs.columbia.edu
4: Princeton University, Princeton, NJ, USA – fellbaum@princeton.edu

Abstract

We analyze how different conceptions of lexical semantics affect sense annotations and how multiple sense inventories can be compared empirically, based on annotated text. Our study focuses on the MASC project, where data has been annotated using WordNet sense identifiers on the one hand, and FrameNet lexical units on the other. This allows us to compare the sense inventories of these lexical resources empirically rather than just theoretically, based on their glosses, leading to new insights. In particular, we compute contingency matrices and develop a novel measure, the Expected Jaccard Index, that quantifies the agreement between annotations of the same data based on two different resources even when they have different sets of categories.

Keywords: lexical semantics, statistical methods, lexical resources

1. Introduction

One of the persistent issues in the use of lexical resources is alignment, both between different lexical resources, and between different versions of the same resource. In this paper, we discuss a study on aligning two of the most widely used lexical resources for English, WordNet (Fellbaum, 1998) and FrameNet (Fillmore and Baker, 2010). There have been a number of attempts to align word senses between these two resources, based on the structure of the databases themselves. In this paper, we present the first empirical comparison of WordNet and FrameNet based on expert sense annotations of real-world text. We found that existing statistics were inadequate for such a comparison, and have devised a new statistical measure, the **Expected Jaccard Index** for this purpose.

1.1. WordNet and FrameNet

WordNet (hereafter WN) is more than ten times larger than FrameNet (hereafter FN), and the two resources have different theoretical bases. In WN, a word sense is represented by including the lemma in a set of near-synonyms (synset); a polysemous word will appear in several synsets. The gloss of a synset is taken as the definition of all the words in it. The synsets are then related to each other via relations such as hyponymy/hypernymy, sister terms, meronymy, etc. Each synset includes only words of one part of speech.

In FN, a word sense (called a **lexical unit or LU**) is represented by including the lemma in a semantic frame and giving it a definition[1]; frames are defined purely semantically, and represent a type of event, relation, state, or entity. Words of different parts of speech can be included in the same frame; there are separate definitions for each frame and for each lexical unit in the frame, and there is no assertion of synonymy among members of a frame–in fact, antonyms are often included in the same frame. The frames are linked by a set of seven semantic relations, such as Inheritance, Subframe (subtype), Causative_of, etc.; there are also sets of frame element to frame element relations corresponding to each of these frame-frame relations. For the majority of the lexical units, FrameNet also includes roughly 10-15 annotated sentences, with labels showing the semantic roles of constituents in the sentence vis-à-vis the frame evoking word or expression.

The two resources are in many ways complementary, with WN providing paradigmatic information about words grouped on the basis of synonymy, and FN providing syntagmatic and valence information about words grouped by the type of event, relation, etc. they describe (Baker and Fellbaum, 2009; Fellbaum and Baker, Forthcoming). Many researchers have used them in combination, e.g. Shi and Mihalcea (2005), Coppola et al. (2009), and there have been several efforts to automatically create a large-scale mapping between FrameNet's lexical units and WordNet's word senses, e.g. Ferrández et al. (2010), Palmer (2009), Laparra et al. (2010), and Tonelli and Pianta (2009).

1.2. Annotation

The authors of the present paper have been involved in a recent project that is producing multiple layers of annotation on a selected subset of the American National Corpus (Ide and Suderman, 2004), known as the Manually-Annotated SubCorpus or MASC for short (Ide et al., 2010), which is also discussed in another paper at this conference (Passonneau et al., 2012). Roughly 100 lemmas have been selected for this project, most of them high frequency words of moderate polysemy (about 4-8 senses each). Since two of the layers are annotation with WN senses and annotation with FN LUs and frame elements (FEs), this is clearly an opportunity to compare FN LUs and WN senses on actual corpus examples of usage; with roughly 90-100 randomly selected sentences per lemma, the more common senses will be documented in dozens of examples, and patterns of overlap between WN and FN senses should be detectable.

The annotation for the two resources was carried out independently by separate teams of trained annotators, WN annotation at Vassar and Columbia U, and FN, at ICSI. In the

[1] Occasionally, the same lemma will occur twice in the same frame, e.g. in the Import_export frame, we have both *export_*((act)).n (*the export of these goods*) and *export_*((entity)).n (*these exports are costly*).

WN process, a sample of 50 sentences was annotated with existing WN senses, then any problems that were noted were discussed in consultation with Christiane Fellbaum at Princeton University. Where the corpus data indicated that the WN inventory should be modified, changes were made, and the annotation proceeded with the revised inventory. In the case of FN, the FN inventory of LUs was expanded in the course of annotation, through consultation between the annotators and other FN staff; in most cases, several new LUs were added. The WN inventory was considered in the FN process, but there was no attempt to match it exactly; in fact, some of the more rarer senses in WN were deliberately omitted, provided that they did not occur in the data.

The ANC team at Vassar annotated roughly 1,000 instances of each lemma (provided that there were that many in the American National Corpus) with the WN senses; a subset of 100 were annotated by 3 or 4 people, to provide data for measures of inter-annotator agreement (IAA) as reported in Passonneau et al. (2012) (this conference), and these cross-validation sets were then given to the FN group for annotation.

2. Contingency Tables

The annotated data for both types of annotation can be summarized in contingency tables. A contingency table is a matrix that captures the relationship between two categorical variables. The rows correspond to the possible values of one variable and the columns correspond to the values of the other. The values reflect how often a particular combination of two values of the two variables occurred.

2.1. Contingency Tables for Sense Annotations

In our case, the rows correspond to the possible WN senses of a word and the columns correspond to the possible FN senses of that word. The values in the table represent the number of sentences assigned to a particular combination of WN sense and FN sense, as in Table 2 on the following page.

The full set of contingency tables for the MASC data are browsable at `http://www.icsi.berkeley.edu/~demelo/masc/`; note that the numbers in the cells are hyperlinks to the annotated sentences, so that it is possible to browse the source of the data. In these tables, the rows have been sorted in order of decreasing row totals and the columns in order of decreasing column totals, so the most frequent senses for both resources tend to appear toward the top left of the table.

Since the FrameNet annotations were done by a single person while the WN annotations were done by three or four people, the WN frequencies have been normalized to sum up to 1 per sentence, allowing us to directly compare the two distributions. This is the reason for the fractional values in the tables; when WN annotators disagree about the senses of a sentence, that sentence is essentially split up between the relevant senses.

2.2. Formal Description

More formally, let S be a set of sentences for some target word. Given a sentence $s \in S$, let $F(s)$ be the corresponding set of FN LU annotations[2] (possibly $\emptyset$ if no LU annotation was performed). Additionally, let $W_a(s)$ be the set of WN annotations by annotator $a \in A$. Let $w_1, \ldots, w_m$ be some arbitrary ordering of WN synset IDs assigned to S by any of the 4 WN annotators (i.e. synsets in the set $\bigcup_{s\in S} \bigcup_{a\in A} W_a(s)$) and let $f_1, \ldots, f_n$ be some arbitrary ordering of FN LUs assigned to S (i.e. LUs in the set $\bigcup_{s\in S} F(s)$). The contingency table $\mathbf{C}$ is defined as an $m \times n$ matrix with entries

$$c_{ij} = \sum_{s\in S} \sum_{a\in A} \frac{|W_a(s) \cap \{w_i\}|}{|W_a(s)|} \cdot \frac{|F(s) \cap \{f_j\}|}{|F(s)|}$$

(where 0 divided by 0 is treated as 0).

In practice, we always have $|F(s)| \leq 1$ on our data set, but in a few cases we observed $|W_a(s)| > 1$, when a WN annotator felt that multiple senses may be appropriate.

3. Examples of specific lemmas

3.1. *curious*.a

We begin with a simple example, the adjective *curious*, where the number of senses is low and the agreement between WordNet and FrameNet is rather clear. (The full set of definitions for both WN and FN are given in the Appendix.)

When the FN team began looking at the WN and FN definitions for *curious*, FN had only one LU for this lemma in the Typicality frame (e.g *a curious symbol*). In looking at corpus examples and the three WN definitions we noticed that some of the uses seemed to refer to a permanent characteristic of an individual (*a curious child*) and others to a temporary state (*a trap door that made me curious*), i.e the "individual level" vs. "stage level" distinction (Kratzer, 1995). The pre-existing frame Mental_property seemed suitable for the permanent characteristic sense, but in order to handle the temporary state sense, a new frame was defined, with the awkward name Mental_state_exper_focus, and another LU was created there, giving FN a total of three senses, and a table of correspondence was worked out with the three WN senses.

No.	Frame	WN No.
F1:	Typicality	1
F2:	Mental_stimulus_exper_focus	(3 →) 2
F3:	Mental_property	2

Table 1: Predicted correspondences between FN and WN senses for *curious*.a

At roughly the same time, annotators working with the WN senses at Vassar and Columbia concluded that WN2 and WN3 were difficult to distinguish, and they were collapsed into one. (WN1 was unchanged.) Before beginning the annotation, the FN annotators drew up a simple table of correspondences between WN and FN senses, shown in the right column of Table 1, with the arrow indicating the merging of senses.

[2] For purposes of computing and display, the LUs and WN senses are represented by their ID numbers.

	FN1	FN2	FN3
WN1	∗48.00	0.00	0.00
WN2	0.00	∗17.00	2.00

Table 2: WN-FN Contingency matrix for *curious*.j

Table 2 is the contingency table resulting from the actual annotation for WN and FN senses; the correspondences predicted before the actual annotation of the example sentences are marked with ∗ in this table. In the event, the correspondence is very straightforward. There were only two instances of the Mental_property sense (indicated by the 2 in the third column) from the following sentences:

1. ...they return home to the inevitable questions from curious friends
2. ...Rick mostly keeps things to himself, but is interrogated by curious children.) [W1,W2,W

3.2. *state.n*

No.	Frame and LU name	WN No.
FN1:	Political_locales.state_(internal).n	1
FN2:	Political_locales.state_(sovereign).n	7
FN3:	Leadership.state.n	3, 4
FN4:	State_of_entity.state.n	2, 6
FN5:	Thermodynamic_phase.state.n	5

Table 3: Predicted correspondences between FN and WN senses for *state*.n

Now let us consider the noun lemma *state*.n., which has eight senses in WN, while FrameNet initially had two senses. (Again, the full set of definitions are in the Appendix.) After studying the WN senses and the data, FN wound up with five senses. One of those involved splitting a pre-existing sense with a disjunctive definition,

> COD: a nation or territory considered as an organized political community under one government, or an organized political community or area forming part of a federal republic.

into two senses of the same lemma in the same frame, one for sovereign states, and one for the internal parts of a republic. WN8 is the sense of *State* meaning the U.S. Department of State, which can occur either in the MWE *Department of State* or alone, as an abbreviation of the MWE; FN elected not to create an LU for this fairly rare use.

Once again, the FN annotators drew up a table of expected correspondences between WN and FN senses, shown in Table 3. As frequently happens with other lemmas, FN collapses some WN senses: WN senses 3 and 4 (which can be analyzed as metonymy or coercion or a regular lexical rule), and also WN 2 and 6, i.e. FN regards the state of a person's mind as only a special case of the state of any entity.

Table 4 is the contingency table resulting from the annotation process; as before, the correspondences predicted before the annotation began are marked with ∗.

	FN1	FN2	FN3	FN4	N/A
WN1	∗ 16.67	0.00	0.50	0.00	1.00
WN4	0.83	6.00	∗ 4.75	0.00	6.58
WN8	0.00	0.00	0.00	0.00	(∗) 11.25
WN2	0.00	0.00	0.00	∗ 4.75	1.00
WN3	2.50	0.00	∗ 0.75	0.00	0.17
WN6	0.00	0.00	0.00	∗ 0.25	0.00
N/A	0.00	0.00	0.00	0.00	43.00

Table 4: Contingency Table for annotations of senses of *state*.n

From this table we can see that the strongest correspondence, between FN1 and WN1, was correctly predicted a priori as was that between FN4 and WN2, and to a lesser extent WN6. The predicted correspondence between FN2 and WN7 does not occur, because no sentences were annotated with WN 7; instead, all FN2 instances were marked as WN4, which was not predicted. Also, WN8 (*Dept. of State*) corresponds to sentences which were not annotated by FN. Thus, we see that the experts were able to predict some (but not all) of the contingency in the annotation on the basis of the WN and FN definitions, synsets, and glosses; this is essentially what programs for automatic alignment of WordNet and FrameNet attempt to do.

The outliers are also instructive: For instance, the FN1/WN3 cell (which was not predicted to occur) contains a number of examples that clearly refer to states of the US, which should presumably be WN1 (*constituent administrative district*) rather than WN3 (*government of a sovereign state*); we hypothesize that the example in the definition of WN3, *the state has lowered its income tax*, may have misled the WN annotators, since both the Federal government and many states levy income taxes in the US.

Beyond verbally describing the patterns of the senses in the contingency tables, however, we would like to have a measure of how coherent such matrices are, to be able to recognize places where a good correspondence across resources can be found.

4. Automatic Evaluation of Annotation Correspondence

Contingency tables for different words vary with respect to the degree of agreement that can be observed. Quantifying the agreement between WN and FN senses for a particular word based on a contingency table is unfortunately not as straightforward as one might expect. Intuitively, a perfect alignment obviously is one that gives us an unambiguous one-to-one mapping between two resources. Hence, a contingency table that can be reordered in the form of a diagonal matrix with at least some strictly positive values should obtain a high score. In contrast, when a sense annotation with respect to one resource co-occurs with many different sense annotations in the other resource, the alignment is a lot less clear. Hence, matrices with uniform element values that do not show any clear correspondence between WN and FN should have low scores.

4.1. Desiderata

The measure we are looking for should have the following properties.

- Maximal score for diagonal matrices: A contingency table that can be rearraned in the form of a diagonal matrix with strictly positive values in the diagonal should yield a maximal score (or at least a very high score).
- Low scores for uniform matrices: An m-by-n matrix whose elements all have the same value should yield a score of at most $\min(\frac{1}{m}, \frac{1}{n})$.
- Ability to discriminate between different degrees of association in m-by-1 and 1-by-n matrices: As will be explained later on in greater detail, even if there is just a single FN LU, there can be different degrees of association between the WN senses and that LU.
- Invariance with respect to transposition: It should not matter which annotation scheme we place on which axis.

4.2. The Expected Jaccard Index

In order to assess the contingency tables in accordance with these intuitions, we derived a new measure, the Expected Jaccard Index $\hat{J}(\mathbf{C})$, as

$$\sum_i \sum_j \frac{c_{i,j}^2}{\left(\sum\limits_{i'} \sum\limits_{j'} c_{i',j'}\right) \left[\left(\sum\limits_{i'} c_{i',j} + \sum\limits_{j'} c_{i,j'}\right) - c_{i,j}\right]}.$$

Motivation

This measure has several desirable properties. First of all, it has a probabilistic interpretation relating it to the standard Jaccard similarity coefficient

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

for similarities between sets of items. If we assume for a moment that every sentence is assigned only one single annotation in each of the annotation schemes, then we can define $C_{i,j}$ as the corresponding sets of sentences associated with each matrix element $c_{i,j}$. In this case, it is easy to show that $\hat{J}(\mathbf{C})$ is equal to the expected Jaccard similarity between the set $C_{i,*}$ of sentences in the i-th row and the set $C_{j,*}$ of sentences in the j-th column.

Proof. By definition,

$$E[J(C_{i,*}, C_{*,j})] = E\left[\frac{|C_{i,*} \cap C_{*,j}|}{|C_{i,*} \cup C_{*,j}|}\right].$$

Since each sentence is only assigned a single cell, this is equivalent to

$$\sum_i \sum_j P_{i,j} \cdot \frac{c_{i,j}}{\left(\sum\limits_{i'} c_{i',j} + \sum\limits_{j'} c_{i,j'}\right) - c_{i,j}}$$

$$= \sum_i \sum_j \frac{c_{i,j}}{\sum\limits_{i'} \sum\limits_{j'} c_{i',j'}} \cdot \frac{c_{i,j}}{\left(\sum\limits_{i'} c_{i',j} + \sum\limits_{j'} c_{i,j'}\right) - c_{i,j}}$$

□

Jaccard similarities are defined for standard Zermelo-Fraenkel sets, where set membership is a binary all-or-nothing notion. Our measure generalizes this to fractional values, which means that a sentence can be divided up between different senses when annotators disagree.

Additional Properties

Another useful property is that scores are guaranteed to lie in $[0,1]$ if the input matrix contains only non-negative values, as for the standard Jaccard coefficient.

Additionally, diagonal matrices with strictly positive values are guaranteed to obtain a score of 1.

Proof. Since all non-diagonal matrix entries are zero, $\hat{J}(\mathbf{C})$ simplifies to

$$\sum_i \frac{c_{i,i}^2}{\left(\sum\limits_{i'} c_{i',i'}\right) c_{i,i}} = \frac{\sum\limits_i c_{i,i}}{\sum\limits_{i'} c_{i',i'}} = 1$$

□

The scores of uniform matrices of ones (or non-negative scalar multiples of such matrices) tend towards zero as the matrix size grows. The second desired property is fulfilled.

Proof. If all $c_{i,j}$ in the m-by-n ($m > 0$, $n > 0$) matrix C are equal to some constant $c > 0$, then $\hat{J}(\mathbf{C})$ becomes

$$\sum_i \sum_j \frac{c^2}{mnc\,(mc + nc - c)} = \frac{c}{mc + nc - c}$$

$$= \frac{1}{m + n - 1} \leq \min(\frac{1}{m}, \frac{1}{n})$$

□

The measure can be applied to m-by-1 matrices like the one presented in Table 5. We will see below that some other measures do not produce appropriate scores for such a table. It is also trivial to prove that the Expected Jaccard is invariant to transposition, because rows and columns are treated alike.

One property that the Expected Jaccard Index does not possess is invariance with respect to lumping and splitting. There are cases where one resource lumps together closely related uses of a word, while another resource splits them into separate entries. The Expected Jaccard Index detects such cases, treating them simply as disagreements between the resources. In some circumstances, however, given that lumping vs. splitting is a relatively well-understood phenomenon, one could also wish to investigate alternative measures that **are** invariant with respect to such splits, i.e. measures that ignore disagreements due to the differing granularity of sense divisions in different lexical resources (cf. Palmer et al. (2007)).

	FN1
WN1	83.5
WN2	5.5
WN3	4.5
WN4	0.5

Table 5: WN-FN Contingency matrix for *people*.n

It must further be pointed out that the Expected Jaccard Index is not a statistical significance test. A diagonal matrix with small values will obtain the same perfect score of 1.0 as a diagonal matrix with large values, because in both cases there is a perfect alignment of the annotations. Our measure does not attempt to assess whether this perfect alignment might have happened by chance. It simply reports that a perfect alignment did occur. Thus one must independently ensure that the sample size, in this case the number of annotated sentences, is large enough to draw reasonable conclusions from the score.

4.3. Comparison with other Measures

Phi coefficient

The Phi coefficient (sometimes referred to as the mean square contingency coefficient or as Yule ϕ) measures the correlation of two variables based on how much the contingency table diverges from the diagonal line. However, the Phi coefficient is defined only for dichotomous variables, i.e. 2x2 contingency tables. The Phi coefficient is also related to Pearson's product-moment correlation coefficient, which is defined for numerical variables, not for the categorical variables (sense assignments) we are dealing with.

Chi Square

The χ^2 statistic is a well-known method for assessing contingency tables of nominal data. The statistic compares the observed values with the expected values under a null hypothesis. For the χ^2 test of independence, this null hypothesis is the hypothesis that the two dimensions are statistically independent.

Unfortunately, the expected values for this test are undefined if any of the rows or columns are empty, as is often the case with our matrices. Additionally, the expected values exactly match the sample values when there is only one row, so the χ^2 score becomes zero, although semantically there might be a high correspondence between the senses, as in Table 5 (*people*.n), where the vast majority of annotations match very well (WN1-FN1).

Sometimes, the χ^2 statistic is instead used with a uniform distribution as the null hypothesis. In that case, however, words like *curious*.j, which have almost perfectly matching annotations (Table 2), score significantly lower than words like *find*.v (Table 8), where an alignment of the annotations was found to be far from obvious.

Mutual Information

Mutual Information (MI) is an information-theoretic measure that reflects how much knowing one variable's value reduces the uncertainty about the other variable. One might expect that such a measure will reveal how much knowing a WN sense reveals about a FN sense or vice versa. In other words, if, for a given token, knowing what sense it is in resource A makes it easy to predict what sense it will be in resource B, then the sense divisions in A and B must be very similar.

Word	Expected Jaccard
curious.j	0.947
high.j	0.902
board.n	0.877
Greek.j	0.860
strike.n	0.856
number.n	0.807
people.n	0.795
—-	—-
state.n	0.630
—-	—-
level.n	0.350
familiar.j	0.312
control.n	0.292
trace.n	0.257
show.v	0.232
find.v	0.212

Table 6: Examples of Expected Jaccard Scores for MASC words

Unfortunately, mutual information scores do not properly handle tables with only a single row or only a single column as in Table 5. From a mutual information perspective, all WN synsets co-occur only with FN1, so knowing the FN LU does not reduce the uncertainty about WN synsets. However, the distribution is highly skewed, and as we see in in Table 5, the first WN sense occurs much more frequently with the FN LU than any of the alternatives, so in reality the annotations are quite similar.

Cramér's V

Cramér's V is a well-known measure of association for nominal data. It is computed by dividing the χ^2 statistic by the sample size and the length of the minimum dimension minus 1, and subsequently taking the square root. Unfortunately, Cramér's V, too, fails to capture the correspondence between the sense schemes shown in Table 5. We merely obtain a value of zero (no association).

Goodman and Kruskal's Lambda

Goodman and Kruskal's λ determines the degree of error reduction one can expect when predicting the value of one variable when one is additionally given the value of a second categorical variable. This measure is well-motivated in certain scenarios, but when comparing sense annotations it does not reflect our intuitions about the example given in Table 5. The additional knowledge of the second variable does not change the prediction about the first, although there is a significant overlap in the annotations.

4.4. Discussion

Overall, we see that these alternative measures are inadequate for this particular task of comparing sense annotations, while the Expected Jaccard Index has several desirable properties and is able to reflect our intuitions about

	FN1	FN2	FN3
WN1	19.50	2.67	0.00
WN2	1.00	8.92	6.17
WN3	1.50	9.92	1.33
WN4	10.33	1.17	0.50
WN5	10.33	0.00	0.00
WN6	3.00	0.00	0.00
WN7	6.33	0.00	0.00
WN8	2.00	0.33	0.00
WN9	0.00	1.00	0.00
WN10	0.00	1.00	0.00

Table 7: WN-FN Contingency matrix for *show*.v

	FN1	FN2	FN3
WN1	39.00	1.67	0.00
WN2	2.00	24.33	0.00
WN3	0.00	0.00	5.00

Table 9: WN-FN Contingency matrix for *strike*.n

good sense alignments. Examples of high, mid, and low scores are shown in Table 6. The measure distinguishes words with sense definitions that align well from words that do not align well. In the latter case, the contingency tables allow us to investigate individual pairs of senses and the corresponding sets of sentences.

Table 7 provides an example of a low-scoring word with word senses that do not line up very well. Depending on the sentence, FrameNet's LU in the Cause_to_perceive frame (FN1) may correspond to WN1 (showing in or as in a picture), WN4 (make visible), WN5 (give an exhibition to an audience), etc. The second FrameNet sense FN2, which is associated with the Evidence frame, also lacks a clear match in WordNet.

For the word *find*, shown in Table 8 on the next page, we obtained the lowest ranking out of all currently annotated words. There are a large number of senses, and for many of them, the alignments far from clear.

A word with a good Expected Jaccard score is shown in Table 9, where we clearly see the alignment along the diagonal, similar to that found in the discussion of *curious* above.

Some other observations can be drawn from the data. There is some tendency for verbs to have lower EJIs, but in general, the part-of-speech of a word is not a good indicator of how well the senses are going to align. Furthermore, lemmas with more senses do not necessarily have worse alignments in the contingency tables. For instance, one of the highest-ranked words, the noun *board*, has 6 relevant WN senses and 6 relevant FN senses. In contrast, the noun *trace* has only 2 relevant WN senses and 5 relevant FN senses, but is one of the lowest-scoring words. [3] This result is in line with the findings in a recent study, carried out on the MASC WN annotations, which showed that a greater number of senses does not necessarily entail lower inter-annotator agreement scores (Passonneau et al., 2010).

Dataset	Pearson r
MapNet v0.1 (Tonelli and Pighin, 2009)	0.259
WordFrameNet (Laparra et al., 2010)	0.350
Ferrández et al. (2010)	0.273

Table 10: WN-FN Contingency matrix for *people.n*

Finally, we performed a comparision between the values found in the contingency tables from the annotation and earlier efforts at WN-FN alignment (at least the three for which we were able to get detailed data). In this computation, the results from Tonelli and Pighin (2009) and Laparra et al. (2010) were binary (i.e. aligned or not aligned), while those from Ferrández et al. (2010) were continuous. For the annotation data, we computed a fractional version of the Jaccard coefficient for each WN-FN sense pair for every lemma, i.e.

$$\frac{c_{i,j}}{\left(\sum_{i'} c_{i',j} + \sum_{j'} c_{i,j'}\right) - c_{i,j}}.$$

We eliminated all WN-FN senses pairs for which the denominator of this fraction is < 5. Table 10 shows the the Pearson correlations of this fractional Jaccard coefficient with the alignments given by other resources.

Due to differences between versions of WordNet and FrameNet and changes made to these resources during the course of the MASC annotation, the resulting scores are not entirely comparable. Still, the relatively low correlations suggest that gloss similarities and other features typically used by automatic methods sometimes fail to reflect the semantic proximity or distance between senses that can empirically be observed when human annotators provide sense annotations for relevant textual data.

5. Conclusions

- We have just begun to analyze the multiply annotated MASC corpus data to compare the WordNet and FrameNet annotations, but we can already see how it can help us single out outliers, which represent sentences which do not fit well in one framework or the other (or either!).
- We have generated contingency tables that accurately summarize the alignment between WN and FN annotations, given different sense inventories and different numbers of annotators.
- We have defined a new measure to assess the correspondence of annotations reflected in contingency tables for nominal data with many zeroes and fractional values. This measure reveals useful information about words and their sense annotations.

6. References

Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation*

[3] This may be because WN and FN use different sense definition criteria for this word. For example, the WN sense *an indication that something has been present* can, among other things, correspond to an ingredient, to a quantified mass, or to a graph shape (e.g. EEG traces) in FN.

	FN1	FN2	FN3	FN4	FN5	FN6	FN7	FN8	FN9
WN1	19.12	0.75	0.25	1.33	0.00	0.00	0.00	0.25	0.00
WN2	10.38	6.12	0.25	2.08	0.25	0.00	0.00	0.00	0.00
WN3	5.00	3.00	2.50	0.25	1.25	0.00	0.00	0.25	0.25
WN4	0.75	0.25	5.00	0.00	1.25	0.00	0.00	0.00	0.50
WN5	2.50	2.00	0.50	0.00	1.25	0.00	0.00	0.00	0.00
WN6	4.50	0.00	0.25	0.00	0.00	0.00	0.00	0.25	0.00
WN7	0.67	0.75	0.00	0.00	0.00	2.75	0.00	0.00	0.00
WN8	1.50	1.62	0.00	1.08	0.00	0.25	0.00	0.00	0.00
WN9	0.50	0.00	2.25	0.25	0.00	0.00	0.00	1.25	0.00
WN10	2.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
WN11	0.00	0.25	0.00	0.00	0.00	0.00	2.00	0.00	0.00
WN12	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
WN13	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25
WN14	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 8: WN-FN Contingency matrix for *find*.v

Workshop, pages 125–129, Suntec, Singapore, August. Association for Computational Linguistics.

Bonaventura Coppola, Aldo Gangemi, Alfio Massimiliano Gliozzo, Davide Picca, and Valentina Presutti. 2009. Frame detection over the semantic web. In *Proceedings of the 6th European Semantic Web Conference*.

Christiane Fellbaum and Collin Baker. Forthcoming. Representing verb meaning in complementary resources. *Linguistics*.

Christiane Fellbaum and Collin F. Baker. (in press). Representing verb meaning in complementary resources. *Linguistics*.

Christane Fellbaum, editor. 1998. *WordNet. An electronic lexical database*. MIT Press, Cambridge/Mass.

Óscar Ferrández, Michael Ellsworth, Rafael Muñoz, and Collin F. Baker. 2010. Aligning FrameNet and WordNet based on semantic neighborhoods. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 310–314, Valletta, Malta, may. European Language Resources Association (ELRA).

Charles J. Fillmore and Collin F. Baker. 2010. A frame approach to semantic analysis. In Bernd Heine and Heiko Narrog, editors, *Oxford Handbook of Linguistic Analysis*, pages 313–341. OUP.

Nancy Ide and Keith Suderman. 2004. The american national corpus first release. In *Proceedings of the Fourth Language Resources and Evaluation Conference(LREC)*, pages 1681–84, Lisbon.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca J. Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Uppsala, Sweden, July. Association for Computational Linguistics.

Angelika Kratzer. 1995. Stage level and individual level predicates. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*. The University of Chicago Press, Chicago.

Egoitz Laparra, German Rigau, and M. Cuadros. 2010. Exploring the integration of WordNet and FrameNet. In *Proceedings of the 5th Global WordNet Conference (GWC'10)*, Mumbai, Jan.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

Martha. Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon ConferenceGenLex-09*, Pisa, Italy, Sept.

Rebecca J. Passonneau, Ansaf Salleb-Aoussi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Rebecca J. Passonneau, Christiane Fellbaum, Nancy Ide, and Collin F. Baker. 2012. The masc word sense sentence corpus. In *Proceedings of LREC 2012*. ELRA.

Lei Shi and Rada Mihalcea. 2005. Putting the pieces together: Combining FrameNet, VerbNet, and WordNet for robust semantic parsing. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico.

Sara Tonelli and Emanuele Pianta. 2009. A novel approach to mapping FrameNet lexical units to WordNet synsets. In *Proceedings of IWCS-8*, Tilburg, The Netherlands, January.

Sara Tonelli and Daniele Pighin. 2009. New features for FrameNet - WordNet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 219–227, Boulder, Colorado, June. Association for Computational Linguistics.

Appendix

A Senses of *curious*.a

WordNet synsets (before study)

1. Synset: *curious, funny, odd, peculiar, queer, rum, rummy, singular* (beyond or deviating from the usual or expected) ”a curious hybrid accent”; ”her speech has a funny twang”; ”they have some funny ideas about war”; ”had an odd name”; ”the peculiar aromatic odor of cloves”; ”something definitely queer about this town”; ”what a rum fellow”; ”singular behavior”
2. Synset: *curious* (eager to investigate and learn or learn more (sometimes about others’ concerns)) ”a curious child is a teacher’s delight”; ”a trap door that made me curious”; ”curious investigators”; ”traffic was slowed by curious rubberneckers”; ”curious about the neighbor’s doings”
3. Synset: *curious* (having curiosity aroused; eagerly interested in learning more) ”a trap door that made me curious”

FrameNet LUs (after study)

1. Typicality.*curious*.a: FN: unorthodox or unexpected
2. Mental_stimulus_exp_focus.*curious*.a: FN: interested or inquisitive (about something)
3. Mental_property.*curious*.a: FN: driven to investigate and learn

B Senses of *state*.n

WordNet synsets

1. Synset: {state, province} Gloss: the territory occupied by one of the constituent administrative districts of a nation; ”his state is in the deep south”
2. Synset: {state} Gloss: the way something is with respect to its main attributes; ”the current state of knowledge”; ”his state of health”; ”in a weak financial state”
3. Synset: {state} Gloss: the group of people comprising the government of a sovereign state; ”the state has lowered its income tax”
4. Synset: {state, nation, country, land, commonwealth, res publica, body politic} Gloss: a politically organized body of people under a single government; ”the state has elected a new president”; ”African nations”; ”students who had come to the nation’s capitol”; ”the country’s largest manufacturer”; ”an industrialized land”
5. Synset: {state of matter, state} Gloss: (chemistry) the three traditional states of matter are solids (fixed shape and volume) and liquids (fixed volume and shaped by the container) and gases (filling the container); ”the solid state of water is called ice”
6. Synset: {state} Gloss: a state of depression or agitation; ”he was in such a state you just couldn’t reason with him”
7. Synset: {country, state, land} Gloss: the territory occupied by a nation; ”he returned to the land of his birth”; ”he visited several European countries”
8. Synset: {Department of State, United States Department of State, State Department, State, DoS} Gloss: the federal department in the United States that sets and maintains foreign policies; ”the Department of State was created in 1789”

FrameNet LUs (after study)

1. Political_locales.state_(internal).n: FN: an organized political community or area forming part of a federal republic.
2. Political_locales.state_(sovereign).n: FN: a nation or territory considered as an organized political community under one government.
3. Leadership.state.n: FN: a metonymic term for governmental actions, departments, and personnel
4. State_of_entity.state.n: the condition of someone or something.
5. Thermodynamic_phase.state.n: FN: general descriptive term of a phase of matter.