# MASC: The Manually Annotated Sub-Corpus of American English

**Nancy Ide*, Collin Baker**, Christiane Fellbaum†, Charles Fillmore**, Rebecca Passonneau††**

*Vassar College
Poughkeepsie, New York USA

**International Computer Science Institute
Berkeley, California USA

†Princeton University
Princeton, New Jersey USA

††Columbia University
New York, New York USA

E-mail: ide@cs.vassar.edu, collinb@icsi.berkeley.edu, fellbaum@princeton.edu, fillmore@icsi.berkeley.edu, becky@cs.columbia.edu

### Abstract

To answer the critical need for *sharable, reusable* annotated resources with rich linguistic annotations, we are developing a Manually Annotated Sub-Corpus (MASC) including texts from diverse genres and manual annotations or manually-validated annotations for multiple levels, including WordNet senses and FrameNet frames and frame elements, both of which have become significant resources in the international computational linguistics community. To derive maximal benefit from the semantic information provided by these resources, the MASC will also include manually-validated shallow parses and named entities, which will enable linking WordNet senses and FrameNet frames within the same sentences into more complex semantic structures and, because named entities will often be the role fillers of FrameNet frames, enrich the semantic and pragmatic information derivable from the sub-corpus. All MASC annotations will be published with detailed inter-annotator agreement measures. The MASC and its annotations will be freely downloadable from the ANC website, thus providing maximum accessibility for researchers from around the globe.

## 1. Overview

To answer the critical need for *sharable, reusable* annotated resources with rich linguistic annotations, we are developing a Manually Annotated Sub-Corpus (MASC) including texts from diverse genres and manual annotations or manually-validated annotations for multiple levels, including WordNet senses and FrameNet frames and frame elements, both of which have become significant resources in the international computational linguistics community. To derive maximal benefit from the semantic information provided by these resources, the MASC will also include manually-validated shallow parses and named entities, which will enable linking WordNet senses and FrameNet frames within the same sentences into more complex semantic structures and, because named entities will often be the role fillers of FrameNet frames, enrich the semantic and pragmatic information derivable from the sub-corpus. All MASC annotations will be published with detailed inter-annotator agreement measures.

The MASC consists of unrestricted (public domain) texts drawn from the American National Corpus (ANC). The corpus and its annotations will be freely downloadable from the ANC website, thus providing maximum accessibility for researchers from around the globe. In addition to providing an invaluable resource for NLP research, the MASC project will contribute significantly to the development of best practices for corpus creation, annotation, and harmonization of annotations from diverse sources on both linguistic and representational grounds.

Because the MASC is an open resource that the community can continually enhance with additional annotations and modifications, it will serve as a model for community-wide resource development. Past experience with corpora such as the *Wall Street Journal* shows that the community is eager to annotate available language data, and we can expect even greater interest in MASC, which includes language data covering a range of genres that no existing resource provides. Therefore, we expect that as MASC evolves, more and more annotations will be contributed, and we can move toward distributed development and a merging of independently developed resources to provide a massive, inter-linked linguistic infrastructure for the study and processing of American English in its many genres and varieties. In addition, by virtue of its WordNet and FrameNet annotations, MASC will be linked to parallel wordnets and framenets in languages other than English, thus creating a global resource for multi-lingual technologies, including machine translation.

## 2. MASC composition

Materials in MASC are drawn primarily from the existing 15 million word Open ANC (OANC)[3], which is

[3] http:// AmericanNationalCorpus.org/OANC

free of any licensing restrictions. The OANC includes traditional genres as well as newer genres such as blogs and email, and is annotated for sentence boundaries, part-of-speech (Penn, Biber, CLAWS5 and CLAWS7 tagsets), and noun and verb chunks. In addition to texts from the OANC, MASC will include portions of existing available corpora that have been manually produced or validated by other projects, such as the *WSJ* corpus annotated by the Penn Treebank II and Discourse Treebank, PropBank, NomBank, and TimeBank.

The outermost hexagon in Figure 1 represents the entire ANC, with each next interior hexagon representing a smaller subset of the data and each wedge representing a different genre. Given the issues outlined above, we expect the contents of MASC, relative to the entire ANC, to follow the pattern of the shaded areas. FN annotations, which are time-intensive to produce, will be done for a genre-representative subset of the data manually validated for WN, entities, and shallow parse. Existing genre-specific data with manually produced annotations, such as the *WSJ* and *Slate* co-reference annotation, will be included in the core in proportions equal to other genres (the remainder of that data will also be made available as a part of the ANC itself). Since examples of phenomena not adequately represented in the core may be required for training, small amounts of data from other parts of the ANC will also be manually annotated to serve this purpose.

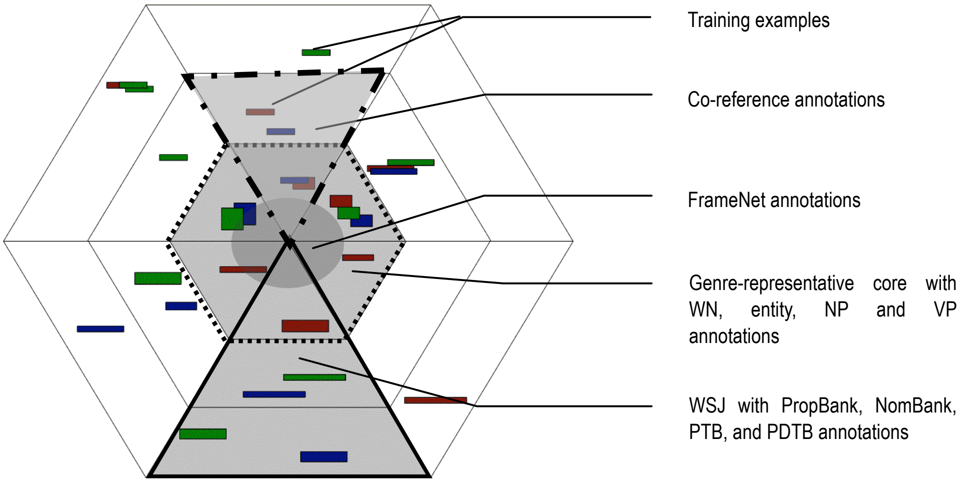
Training examples
Co-reference annotations
FrameNet annotations
Genre-representative core with WN, entity, NP and VP annotations
WSJ with PropBank, NomBank, PTB, and PDTB annotations

Figure 1. Configuration of MASC relative to the entire ANC

The MASC annotation process proceeds as follows: smaller portions of the sub-corpus are first manually annotated for specific phenomena, with an eye toward maintaining representativeness in these smaller portions as well as ensuring that a common component includes as many annotations of different types as possible. We then apply (semi)-automatic annotation techniques to determine the reliability of their results, and study interannotator agreement on manually-produced annotations in order to determine a benchmark of accuracy and fine-tune annotator guidelines. We also consider the degree to which accurate annotations for one phenomenon can improve the performance of automatic annotation systems for another—e.g., validated WN sense tags and noun chunks may improve automatic semantic role labeling. We then apply an iterative process of manual annotation followed by retraining of automatic annotation software to maximize the performance of the automatic taggers. The improved annotation software can later be applied to the entire ANC, thus providing more accurate automatically-produced annotation of this much larger body of data.

## 3. Representation

The ANC project has implemented a scheme for representing the ANC and its annotations that answers one of the field’s urgent needs, that is, means for individual annotations to cohabit with one another and/or be layered over primary data, using a common format that allows them to be easily merged and manipulated (Ide and Suderman 2006, 2007). The perennial problem for language processing research is the fact that annotations produced at different sites are idiosyncratic and often demand considerable processing to render them usable with software for which they were not originally designed, or to combine annotations produced at different sites. For example, the *Wall Street Journal* corpus has been annotated by several groups for different phenomena: The Penn Treebank (PTB) II syntactic annotations are embedded in the data itself in a LISP-like format; PropBank and NomBank reference the data by sentence (tree) number and token number; and the Penn Discourse Treebank uses a complex addressing system to reference nodes in the PTB constituency trees. Using any one of these annotations demands that one’s software can process their addressing mechanisms (which typically requires programming effort to adapt), and merging them is very far from a trivial task. The result is that individual projects and groups spend considerable time and effort to massage annotation information into a format they can use, involving much duplication of effort and time spent on low-level tasks prior to addressing research questions.

The philosophy of the ANC scheme is maximum flexibility. This is accomplished by first rendering all annotations into a generic, feature structure-based format[4] and outputting each in a separate stand-off document linked either to the primary data (which is text-only and read-only) or to other annotations. Annotations in this format utilize the *original annotation labels*—i.e., we make no effort to provide annotations whose content categories are harmonized on linguistic grounds. Users then use a graphical interface to the ANC's freely-distributed ANCTool[5] to select the annotations they are interested in, and the tool produces a version of the corpus with the merged annotations in-line. The output format of the data and individual or merged annotations is controlled by the user, and can therefore be tailored to a particular software system or use. The current version of the ANC tool provides several built-in output options, including XML[6] (suitable for input to the BNC's XAIRA system) and non-XML formats that can be used with systems such as NLTK and concordancing tools. The ANCTool can also produce merged annotations in GrAF format to enable the application of basic graph traversal algorithms to merged annotations or provide input to graph visualization tools such as GraphViz.[7] We are currently adapting the ANCTool to generate corpora and annotations in UIMA format.[8] However, because the tool's underlying parser uses multiple implementations of the org.xml.sax.DocumentHandler interface (one for each output format), additional formats are trivially generated by implementing additional interfaces.

## 4. WordNet annotation of the MASC

There is a large number of state-of-the-art word sense disambiguation (WSD) systems based on WordNet senses, several of which are freely available for research purposes (e.g. Pederson's SenseRelate system, Mihalcea *et al.*'s SenseLearner). Some of these systems were among the top performers in SENSEVAL-3; for example, SenseLearner was second overall in the English-all-words task (Snyder and Palmer, 2004). We are updating these systems to use the most recent WordNet version, 3.0; they will then be applied to automatically assign WN sense tags all content words (nouns, verbs, adjectives, and adverbs) in the entire ANC. The resulting sense annotations are serving as the basis for the manual correction of the MASC, which will include the FrameNet-annotated portion.

The manual sense-tag correction is being performed by a team of undergraduates from several institutions. We are building on the experience from a recent Vassar-Princeton pilot sense tagging project, where student annotators manually assigned WN sense tags to all occurrences of a small selection of nouns and verbs in the ANC 2nd Release data. For this purpose the WN annotation software was modified to generate the sense-tagged ANC data in a form that enables automatic production of stand-off annotation documents containing the annotations. Annotators are trained and provided with a tagging manual, and all annotation is subject to careful quality controls and validation to ensure that annotation policies are consistently followed.

## 5. FrameNet annotation of the MASC

The FrameNet project is developing a lexicon of English based on the theory of Frame Semantics (Fillmore 1976), centered around the concept of semantic frames, each of which represents an event, relation, state, or (occasionally) entity. In FrameNet annotation of texts, each predicator (which may be a verb, noun, adjective, adverb or preposition is labeled with the name of the frame it **evokes**, and arguments (and sometimes adjuncts) of the predicator are labeled according to the role they play in the situation of the frame; these roles are known as **frame elements (FEs)**, and are specific to each frame. The frames and FEs are connected by relations such as inheritance, sub-event, causative-of, etc. (Fillmore et al. 2004, Lönneker-Rodman & Baker ms.)

The FrameNet team is involved in two rather different annotation tasks for the MASC project:

- full manual annotation of a subcorpus (smaller than the MASC) in the usual FrameNet full-text manner (similar to the so-called "all-words" tasks in the Sensevals[9]), and
- application of automatic semantic role labeling software over the whole MASC and providing the results of that automatic labeling to the ANC consortium.

[4] The ANC format is based on ISO TC37 SC4's Linguistic Annotation Framework, which is isomorphic to other feature structure-based representations such as UIMA's Common Analysis Structure.

[5] See http://AmericanNationalCorpus.org/tools.html

[6] For XML output, the user also chooses how to handle overlapping hierarchies from among several options.

[7] http://www.graphviz.org/

[8] Supported by an IBM Innovation Award; this option should be available by the time of the LREC 2008 conference.

[9] Cf. Mihalcea & Edmonds 2004. In fact, of course, in all such tasks, there is always some set of words that are specifically **not** to be annotated.

Figure 2 shows part of the FrameNet annotation of one sentence from the ANC, from a travel guide to Dublin. *The River Liffey flows from west to east through the center of the city to Dublin Bay.* The three rows represent annotation in three different frames. Row 1 represents annotation in the frame Fluidic_motion; The work *flows* evokes the frame. *The River Liffey* is labeled as the FE Fluid, and the Source FE is expressed by *from west*, the Area FE, by *through the center of the city*, and the Goal FE by two separate phrases, *to east* and *to Dublin Bay*.; The FEs Source, Path, Area, and Goal occur in all the frames that inherit from the high-level frame Motion. Row 2 shows the annotation for the frame Part_inner_outer, evoked by the word *center*; *center* itself also denotes the FE Part, and the FE Whole is represented by the PP *of the city*. Row 3 gives two separate instances of the frame Natural_feature, one evoked by *River*, and the other by *Bay*; in each case, the word itself denotes the FE Locale, and the FE Name follows in the one case and precedes in the other. (Note that expressions denoting natural features are idiosyncratic and have to be learned individually; River Liffey but Mississippi River, Dublin Bay, but Bay of Bengal, etc.) It should be clear that by correctly composing the information contained in these annotations (and others not shown for reasons of space), one should be able to make many valid inferences: that there is a river whose name is Liffey which flows through a place which is the inner part of a city, to a bay whose name is Dublin, etc. More elaborate types of inference should also be supported by FrameNet annotation, as discussed in Scheffczyk et al. (2006).

| | |
|---|---|
| 1 | [$_{\text{FLUID}}$ The River Liffey] FLOWS [$_{\text{SOURCE}}$ from west] [$_{\text{GOAL}}$ to east] [$_{\text{AREA}}$ through the center of the city] [$_{\text{GOAL}}$ to Dublin Bay]. |
| 2 | The River Liffey flows from west to east through the [$_{\text{PART}}$ CENTER Target] [$_{\text{WHOLE}}$ of the city] to Dublin Bay. |
| 3 | The [$_{\text{LOCALE}}$ RIVER] [$_{\text{NAME}}$ Liffey] flows from west to east through the center of the city to [$_{\text{NAME}}$ Dublin] [$_{\text{LOCALE}}$ BAY]. |

Figure 2. Example of FrameNet annotation

The automatic labeling of frames is not as well-developed as the WSD algorithms for WN sense assignment, and the job of recognizing FEs adds another task of some complexity. The automatic semantic role labeling systems usually consist of two separate processes, each treated as a classification problem: (1) recognizing which words (or multi-word expressions) evoke which frames, and (2) labeling the arguments of such words with the correct FE (role) labels. Errors in the frame recognizer consist either of assigning a word (or MWE) to the wrong frame or to no frame where the correct frame exists. The frame element labeler (a.k.a. semantic role labeler) can produce a variety of errors, by failing to label FEs where they belong or by labeling the right text with the wrong FEs, or by misidentifying the boundaries of the FE.

Despite the difficulty, and thanks largely to the impetus of two separate competitions concerned with frame semantic annotation, at Senseval-3 (Litkowski et al. 2004) and Semeval-4 (Baker et al. 2006), there are currently three publicly available systems for automatically recognizing frames and assigning the semantic role (frame element) labels:

- Shalmaneser, developed by Sebastian Padó and Katrin Erk at University of Saarland (Erk,& Padó 2006, http://compling.utexas.edu/shalmanesar),
- the ASSERT system, developed at University of Colorado by Sameer Pradhan, (Pradhan et al. 2004, http://cemantix.org/assert) which has been used mainly for PropBank-style role labeling, but has also been trained on the FrameNet data, and
- the system developed by Richard Johansson and Pierre Nugues at Lund University for the most recent Semeval (Johansson & Nugues 2007).

As part of this work, the FrameNet team is also committed to improving the ASRL process, using an active learning system, whereby the ASRL system results are evaluated to determine where the most errors were occurring, and extra manual annotation is done to improve performance and reduce those errors. In some cases, the ASRL systems themselves can output a confidence measure; another approach is to use several systems and to concentrate on those cases in which the different systems disagree.[10] The sentences to be manually annotated for this purpose could come from anywhere, but we plan to use the entire ANC (not just the MASC) for this purpose.[11] The supplemental annotation is close to the usual FrameNet lexicographic annotation in terms of process, although it may involve more (or less) examples per LU than the usual 15-20 for FrameNet. Also, examples are chosen that are close to the boundary of

[10] Note that the three systems use different feature sets, different machine learning algorithms, and even define labeling differently, some labeling nodes in a parse tree and others labeling spans of text.

[11] It may be necessary to reach beyond the ANC to the BNC or the web, although that has the unfortunate consequence that these supplemental annotations would not be able to be included in the ANC.

other senses, rather than central, prototypical uses, as is usual in lexicography.

It is not definite whether the team will be able to improve on the algorithms now used in ASRL systems, but just adding selected additional annotation should produce significantly better output from the current systems. The entire text will be repeatedly automatically labeled as the accuracy of the process improves, and one of the deliverables will be a good, largely automatic annotation of at least the entire MASC (and possibly the whole ANC, if the system runs fast enough).

Note that neither of these tasks is very close to what in the Sensevals is called the "lexical sample" style of task, where a few words are annotated across a large amount of text.

## 6. Alignment of Lexical Resources

A concurrent project is now investigating how and to what extent WordNet and FrameNet can be aligned with each other. As those familiar with both resources will be aware, WordNet is much larger than FrameNet, and their structure is quite different. But they are in many ways complementary, and a mechanism for accessing the information available from both resources in the same way would be useful for many NLP purposes. Since the same text will be annotated both for FrameNet frames and frame elements and (independently) for WordNet senses (synsets) as part of the MASC, this will provide a ready-made testing ground for the WordNet-FameNet alignment. As further annotations from other projects are added to MASC, similar studies for aligning them should become feasible.

## 7. Interannotator agreement

We assess the manual annotations in MASC using a suite of metrics that measure different characteristics. To determine whether annotators agree at a level above chance, we use interannotator agreement coefficients, such as Cohen's kappa (Cohen, 1960. To determine what proportion of the annotated data all annotators agree on, we use average F-measure (van Rijsbergen, 1979). To determine the impact of these two measures of quality, we consider the relation between the agreement coefficient values and F-measure and with potential users of the planned annotations (or where possible, with existing users of annotations that will be contributed to MASC), to determine whether they can provide independent performance measures for applications of the data using data from different annotators. In previous work (Passonneau et al. 2005; Passonneau et al. 2006) we have argued that simultaneous investigations of interannotator agreement, and measurable results of using different annotations of the same data, provide a stronger picture of the integrity of annotated data, given that there are no absolute criteria for what constitutes good interannotator agreement (Krippendorff 1980).

As in (Passonneau et al. 2006), we partition annotation datasets in subsets for purposes of comparative analysis by genre, modality and source. This allows us to assess to some degree, depending on other factors such as whether the same annotators work on all genres, whether annotation quality varies with these factors. Statistics reflecting interannotator agreement levels will be distributed with the MASC data.

## 8. Conclusion

The overall goal for MASC is to continually augment the sub-corpus with contributed annotations from the research community, so that in the future annotations for additional linguistic phenomena such as discourse structure, additional entities, events, opinions, etc. will be added. We feel strongly that distribution of effort, together with integration of currently independent resources such as the ANC, WordNet, and FrameNet, will enable progress in resource development well beyond what can be accomplished at individual sites working independently (which is the model in operation at the moment), for considerably less cost and without duplication of effort, and achieving a greater degree of accuracy and usability. Its availability should have a major impact on the speed with which similar resources can be reliably annotated.

The addition of semantic annotation for WN senses and FN frames will make the MASC the largest semantically annotated corpus of English in existence and provide a much-needed resource for computational linguistics research aimed at the development of robust language processing systems. Because both WN and FN are linked to corresponding resources in other languages, WN and FN annotation of the MASC will immediately create a massive multi-lingual resource network. The unprecedented nature and value of such a resource for machine translation and other multi-lingual NLP applications cannot be underestimated, as no existing resource approaches this scope.

## 9. Acknowledgements

The work described in this paper is supported U.S. National Science Foundation grant NSF CRI 0708952.

## 10. References

Baker, C.; Ellsworth, M. & Erk, K. (2007). SemEval-2007 Task 19: Frame Semantic Structure Extraction *Proceedings of the SemEval 2007 Workshop*, Association for Computational Linguistics.

Cohen, J. 1960. A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20: 37–46.

Erk, K. & Padó, S. (2006). Shalmaneser -- A Flexible Toolbox For Semantic Role Assignment. *Proceedings of the Fifth International Conference* *on Language Resources and Evaluation* (LREC-2006).

Fillmore, C. J. (1976). Frame semantics and the nature of language Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280, 20-32.

Fillmore, C. J.; Baker, C. F. & Sato, H. (2004). Framenet as a "Net" *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC), 1091-1094.

Ide, N. and Suderman, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. *Proceedings of the Linguistic Annotation Workshop*, Association for Computational Linguistics, 1-8.

Johansson, R. & Nugues, P. (2007). LTH: Semantic Structure Extraction using Nonprojective Dependency Trees *Proceedings of the SemEval 2007 Workshop*, Association for Computational Linguistics, 227-230.

Krippendorff, K. (1980.) Content Analysis: An Introduction to its Methodology. Sage Publications, Beverly Hills, CA.

Litkowski, K. Mihalcea, R. & Edmonds, P. (ed.) (2004). Senseval-3 task: Automatic labeling of semantic roles. Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, 9-12.

Lönneker-Rodman, B. & Baker, C. F. The FrameNet Model and its Applications. (ms.)

Mihalcea, R. & Edmonds, P. (ed.) (2004). Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text

Passonneau, R., Habash, N., Rambow, O. 2006. Inter-annotator agreement on a multilingual semantic annotation task. *Proceedings of the Fifith International Conference on Language Resources and Evaluation* (LREC). Genoa, Italy.

Passonneau, R., Nenkova, A., McKeown, K. and Sigelman, S. 2005. Applying the pyramid method in duc 2005. *Proceedings of the 2005 Workshop of the Document Understanding Conference (DUC).*

Pradhan, S. S.; Ward, W. H.; Hacioglu, K.; Martin, J. H. & Jurafsky, D. Susan Dumais, D. M. & Roukos, S. (ed.) (2004). Shallow Semantic Parsing using Support Vector Machines HLT-NAACL 2004: Main Proceedings, Association for Computational Linguistics, 233-240.

Scheffczyk, J.; Baker, C. F. & Narayanan, S. Oltramari, A. (ed.) (2006). Ontology-based Reasoning about Lexical Resources. *Proceedings of ONTOLEX 2006*, 1-8

Snyder, B. and Palmer, M. (2004). The English all-words task. In Mihalcea, R., Edmonds, P. (Eds.), *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Association for Computational Linguistics, 41-43.

van Rijsbergen, C. J. 1979. Information Retrieval. 2nd edition, London: Butterworths.