

Welcome to Release 1.5 of the FrameNet data

Thank you for your interest in FrameNet; we hope that you will find it both interesting and useful in your work. First, a brief explanation of why this release is numbered 1.5: Some time ago, we prepared a FrameNet data set for use as training data in SemEval 2010. Since it was not carefully checked, we called it Release 1.4-alpha, intending to produce a clean version soon afterward. But almost a year has passed since then, and many changes and additions have been made, so we're calling this one Release 1.5. We hope that it will be used from now on instead of the R1.4-alpha data; it should be cleaner and more consistent, as explained below.

1 Major changes in data format

We have drastically revised the format of our data files for this release. We regret the inconvenience this will cause long-time users of the FrameNet data, but there were many compelling reasons to make the change.

The previous data release (Release 1.3) contained both XML and HTML copies of all the data files, so that they would be both machine- and human-readable; we were also producing versions of the XML with and without part of speech (POS) labels. Since there are more than 10,000 lexical units, each with its own file, with 1 HTML and 2 XML files for each, this involved more than 30,000 files, which required a long time to generate, and had to be kept in sync across the different formats. We also produced different versions of the HTML for use on the public website, for internal use, and for the data distribution.

The new system creates just one XML format for all uses; a set of new XSL/Javascript scripts will allow the XML to be viewed by most standard browsers. (See Appendix A for details of browser compatibility.) Even the top-level index files for selecting frames, lexical units, and full texts are XML with accompanying XSL scripts. In order to allow these scripts to run fast enough, we have included redundant information about the colors used to render the FEs in all the files connected with each frame, and also split the information about frames into one file per frame, rather than one large file for all frames. But the total number of files in the distribution has been drastically reduced, along with the overall size.

One advantage is that the process of exporting the release data will be much simpler, so that we should be able to update it more often. Another major advantage is that we have been able to build new functionality into the scripts, so that browsing the data and navigating among the various reports will be fully available to anyone who downloads the data release, in a style very similar to our public website. The user can browse full-text annotation files, clicking on individual predicates to see the annotation connected with them, as on the public website. Furthermore, we have been able to provide a function formerly available only on our internal website: in the valence tables (in the Lexical Entry reports), the numbers showing the counts of sentences falling into each category are themselves links—clicking on them displays the annotated sentences that fall into that category in a separate HTML frame.

We are also moving from DTDs to XML schemas, a more modern system for definition and validation of XML syntax. We hope to produce a set of files documenting differences between R1.3 and R1.5, and tools/APIs for accessing the data in the new format, similar to ones we had comparing R1.2 and R1.3, shortly after this release, although this will take some time, due to the significant changes in data format. (See the XML Documentation file in this directory for further details.)

2 Change to Creative Commons License

FrameNet Release 1.2 was only free to academic researchers. Release 1.3 was made available without charge for research and development purposes to everyone, including for-profit companies. Continuing this trend of less restrictive licenses, we are offering Release 1.5 under a Creative Commons Attribution-Only license. For details of the license, please see the Creative Commons website, <http://creativecommons.org/licenses/by/3.0/>.

ICSI will continue to hold the copyright on the data and documentation so that no one else can change the licensing arrangement. We suspect that a number of firms are using FrameNet as the basis of products and services which they are currently providing. We hope that a shift to a very open license will encourage them to download the new version and to acknowledge their connection to FrameNet and also foster collaborative development of frames in areas of commercial interest. We welcome anyone interested in development of semantic frames in specific domains to contact us to discuss collaboration, and will set up a new website to host such discussions.

3 Growth of the FrameNet database

	R1.2	R1.3	R1.5	Change R1.3→R1.5
Frames	609	795	1019	28%
(non-lexical)	58	74	111	50%
FEs in lexical frames	4909	7124	8884	25%
FE/lexical frame	8.91	9.88	9.78	-1%
Pct. non-lexical	9.5%	9.3%	10.9%	17%
Frame relations	550	1152	1507	31%
FE relations	2770	6311	8252	18%
Lexical Units	8869	10195	11829	16%
LUs/lexical frame	16.1	14.14	13.03	-8%
LUs w/ lexicog anno	6642	6815	7711	13%
Pct. LUs w/ lexicog. anno	74.9%	66.8%	65.2%	-2%
AnnoSets in lexicog anno	133846	139439	149931	8%
Lexicog AnnoSets/annotated LU	20.2	20.5	19.5	-5%
AnnoSets in full text anno	0	11671	23087	98%
Total AnnoSets	133846	151110	173018	14%

Table 1: Some statistics comparing Releases 1.2, 1.3 and 1.5

4 More lexical units, frames, and FEs.

As shown in the summary table above, the number of lexical units has increased by 16% to more than 11,800, while the numbers of frames and FEs have increased by 25% to 1,019 and 8,884 respectively. Some frames are marked as “non-lexical”, meaning that we have created them because they are logically necessary in the frame hierarchy, even though they do not

contain lexical units. The proportion of non-lexical frames has increased from about 9% to about 11%, as the frame hierarchy has continued to be filled out.

The number of LUs per lexical frame has decreased from about 14 to about 13. This may be due in part to the continued, gradual extension of FrameNet to less frequent lemmas with less polysemy and more specialized terms (although we continue to work in many domains simultaneously, so that FN remains as domain-independent as possible). The percentage of LUs with lexicographic annotation has decreased slightly to 65%, primarily because, as we annotate full texts, we need to create new LUs to cover vocabulary found in them, and may not have the resources to keep up the lexicographic annotation at the same pace. For those LUs that do receive regular lexicographic treatment, the average number of annotated sentences per lexical unit has decreased slightly to 19.4; this may also be partially due to the inclusion of less frequent words, for which fewer good examples are available.

Thanks to a subcontract in the MASC project (NSF CRI-CRD 0708952) <http://www.americannationalcorpus.org/MASC/Home.html>) for the annotation of a portion of the American National Corpus with WordNet and FrameNet labels (along with annotations from other sources), we have been able to annotate quite a few new texts in full-text annotation style (i.e. all frame-evoking words are labeled as targets, each with its own annotation set). These are the texts viewable starting from the full-text index, under the ANC corpus heading. Some of these were also used in the SemEval 2007 Task on Frame Semantics. The amount of full-text annotation has almost doubled since Release 1.3; some of the files have only been partially annotated, see Appendix C for a list.

Users of the lexicographic annotation data (i.e. the lexical entry and annotation views) may note that certain LUs have recently been annotated on far more than the 15 to 20 sentences per LU which are usual for lexicographic annotation. This concentrated annotation of small number of highly polysemous lemmas is a result of work on two current projects, one studying the alignment of WordNet and FrameNet (NSF IIS-RI-0705155) and the MASC project.

The number of frame relations and their corresponding frame element relations have increased substantially.¹

5 Improvements in consistency and completeness

As with earlier data releases, we have devoted a great deal of attention to ensuring that the data in this release is consistent and complete in a number of aspects. We recognize that the definitions of frames and FEs are not always sufficient to convey what they are intended to mean, so we have striven to ensure that at least one LU is annotated in each frame, and that there is at least one example of each core FE annotated in each frame. Appendix B lists cases in which we have been unable to accomplish this, with possible reasons.

We have also worked on eliminating incorrect non-ASCII characters in the texts of the sentences, a problem which has been with us for a long time. At a minimum, we believe that

¹The figures shown here for R1.2 and R1.3 are lower than those given in the release notes for Release 1.3, because in this calculation, we have excluded the frame relations in reframing mappings and also the “dummy” frame relations associated with the FE relations Core Set, Requires, and Excludes, counting only the eight frame relations likely to be of use to users of the data: Inheritance, Using, Point of view, Subframe, Precedes, Causative of, Inchoative of, and See also. The FE relations Core Set, Requires, and Excludes are included in the frame.xml for each frame and are displayed when browsing the frames (for the first time in this release).

all the text of the sentences is now valid UTF-8 Unicode. See the XML Documentation for details of this cleanup.

On a more mundane level, we have also tried to eliminate some spelling errors in frame and FE definitions, and to ensure that frame and FE names are consistent both within frame definitions and across frames. We have tried to ensure that LU and lexemes names use American spellings, while including British spellings among the word forms. However, even given these efforts toward consistency, we urge users of our data to depend on the **ID numbers** of frames, FEs and LUs, rather than their names, for comparisons across different versions of the FN data; from time to time, the FN team decides to rename frames, FEs (or occasionally, LUs) and rewrite their definitions to reflect a different (and hopefully clearer) way to carve up the conceptual world, but the ID numbers will remain constant.

The adoption of a single XML representation for each type of data should also mean that the FN public website will be more consistent with the data release.

6 Other representations of FrameNet data

A number of groups outside the Berkeley FrameNet group have created other representations of the FN data, especially with a view toward using it for inferencing, including OWL-DL and Prolog versions of the data. We are glad to cooperate in efforts to create other representations, and would be happy to host or link to them, if appropriate, from the FrameNet public website. We also welcome suggestions on what might be done to make the FN database more amenable to use in inferencing, without sacrificing the accuracy of linguistic description and the corpus basis of our research.

Collin Baker

Jisup Hong 2010.09.09

Appendix

A Browser Compatibility

A.1 Supported web browsers

(This section duplicates a portion of the XML Documentation.)

The following are browser/platform combinations that have been tested and verified to work correctly with the R1.5 XML and XSL/Javascript:

Red Hat EL5:	Firefox 3.0.16, 3.6.7
Ubuntu 9.10:	Firefox 3.5.9; Chrome*
Windows XP SP2:	Firefox 3.6.3; IE 8.0.600**
Windows Vista:	Firefox 3.0.14, 3.6.3; IE 8.0.6001**
Windows 7:	Firefox 3.6.3, IE 8.0.7600; Chrome*
Mac OS X:	Firefox 3.5.2; Safari 4.0.5, 5.0.1; Chrome*

* Chrome seems to work correctly if the command-line switch `--allow-file-access-from-files`

is used; see <http://code.google.com/p/chromium/issues/detail?id=47416> for more information. This problem may be fixed in later versions of Chrome.

******For IE on Windows XP SP2+ and Vista, please see Sec. A.3 below.

A.2 Unsupported web browser(s)

The Opera browser is not supported in Release 1.5. We have tested Opera 10.62 on Mac OS X and 10.60 on RH Linux EL5, and found the same problems in both. Those files that do not involve HTML framesets work perfectly; these include the the full-text index, the LU index and the annotation reports. Thus it is possible to open the LU index (i.e. the file `luIndex.xml`), find an LU and click on the “Annotation” link and see the annotation. But it appears that Opera cannot handle HTML framesets, so the frame index does not work, nor do browsing the full-text annotation files or the lexical entry report.

A.3 Issue with Internet Explorer on Windows XP SP2+ or Windows Vista

When the FN XML files are downloaded and opened with IE on Windows XP SP2 or Vista, users will receive the following error message:

The XML page cannot be displayed: Cannot view XML input using XSL style sheet. Please correct the error and then click the Refresh button, or try again later. | Access is denied.

(This problem does not affect Windows 7 users.) The recommended (and easiest) work-around for users of Windows XP SP2 and Vista is to use Firefox instead of IE.

A detailed explanation of the cause of this problem, and alternate work-arounds are as follows:

In Windows XP SP2+ and Windows Vista, Microsoft implemented a policy whereby files downloaded from the internet (or any any non local source) are marked with an NTFS named data stream called “Zone.Marking”. This stream contains the location from which the file was downloaded. For example, if you download the Google Toolbar Installer, the executable would have the following Zone.Marking stream:

```
[ ZoneTransfer ]
ZoneUri=http://toolbar.google.com/data/
en/big/current/GoogleToolbarInstaller.exe
```

Each time you launch an executable (or script) from Explorer that has this named data stream, either on the same machine or from a network share, the security warning mentioned above is displayed.

There are several options for deleting these streams from files:

1. Delete all streams with the Windows system internal utility called “Streams” (documented at <http://technet.microsoft.com/en-us/sysinternals/bb897440.aspx>) Note that occasionally streams are used for other purposes than zone marking, so deleting all streams could cause other problems.
2. Delete all streams by means of a non NTFS partition/drive. Zone marking is not used outside of NTFS, so copying the files onto a FAT partition (such as on a USB thumb drive) and back will clear the zone marking. (Same caveat as above.)

3. Delete only Zone.Identifier streams using a program such as AlternateStreamView (http://www.nirsoft.net/utils/alternate_data_streams.html). You can delete all streams marked "/:Zone.Identifier:\$DATA/" for the selected files to get rid of the security blocks.
4. Clear the zone marking on each file individually, using the properties->unlock button. This can easily be done for the top-level index.xml and .xsl files, but is hardly practical for all 10,000+ lu XML files.
5. A last option would be to turn off zone marking globally. This is detailed at: <http://www.petri.co.il/unblock-files-windows-vista.htm>

B Frames containing core FEs that are not annotated

As mentioned above, we have made an effort to ensure that every core FE is annotated at least once per frame (and where feasible, once per LU). However, this has not always been possible; some core FEs are still not annotated at all in their frame; there are two types of reasons for this:

(1) The following 12 core FEs in 6 frames are missing annotated evidence simply because the expression of these frame elements is extremely rare in the corpora available to FrameNet:

Frame:	Missing Core Frame Element(s):
Change_position_on_a_scale	FINAL_STATE, INITIAL_STATE ²
Emotions.success_or_failure	EVENT, EXPRESSOR, TOPIC
Exchange_currency	MONEY, SUM_1, SUM_2
Experiencer_focus	EVENT
Reparation	GIFT
Waver_between_options	OPTION_1, OPTION_2

(2) The following 15 core FEs in 8 frames are not necessarily rare, but the frames or FEs need reanalysis or modification; unfortunately, we were not able to do so before the data release:

Frame	Missing Core FE(s):	Notes:
Addiction	STATE	This FE should be denoted by some targets, like <i>alcoholism.n</i> STATE also excludes ADDICT but sentences like <i>His alcoholism is a problem</i> are possible. Frame reanalysis needed.
Frequency	TIME_SPAN	Frame reanalysis needed.
Import Export	AGENT, EX- PORTING AREA, GOODS, IMPORT- ING AREA	Perhaps this frame should be made non-lexical, since FrameNet already has the frames Importation and Exportation, which are in Perspective_on relations to it.
Labor product	CAUSE	Either attestations are very rare or there is a problem with the frame; reanalysis needed.
Mental stimulus Experiencer focus	EXPRESSOR, STATE, TOPIC	Frame reanalysis needed; there are too many frame elements that are too similar to each other.
Performers and roles	AUDIENCE, SCORE, SCRIPT	This frame needs perspectives or, at least, new frames with Point-of-view relations to it (for example, from p.o.v. of audience and from p.o.v. of performer.
Sensation	GROUND	This frame needs revision.
Time_Span	DURATION	Frame reanalysis needed; we should either rephrase definitions of core frame elements or expand the list of Lexical Units.

C Partially annotated full-text files

The following 12 files are listed as full-text annotation, but work on them has not been finished. We decided to include them in the release anyway, in the hope that they may be useful even in their unfinished state. All the other full-text files are more or less “completely annotated”, although, as we add new frames and LUs for other purposes, we continue to revisit these files, adding more targets.

ANC	110CYL067	annotation 50% complete, frame assignment done for the rest of the text (no FE labels added)
ANC	110CYL069	frame assignment done for the whole text (no FE labels added)
ANC	IntroHongKong	annotation 75% complete, rest of text not annotated
ANC KBEval	WhereToHongKong Brandeis	annotation 33% complete annotation 25% complete, 15% of text has frame assignments, and other 60% not annotated
KBEval	cycorp	annotation 10% complete, rest of text not annotated
KBEval	parc	annotation 10% complete
KBEval	Stanford	annotation 20% complete, rest of text not annotated
LUCorpus-v0.3	enron-thread-159550	annotation 50% complete
LUCorpus-v0.3	IZ-060316-01-Trans-1	annotation 50% complete
Miscellaneous	Hound-Ch14	annotation 5% complete
Miscellaneous	SadatAssassination	annotation 80% complete