**NLTK Corpora**

NLTK has built-in support for dozens of corpora and trained models, as listed below. To use these within NLTK we recommend that you use the NLTK corpus downloader, >>> nltk.download()

Please consult the README file included with each corpus for further information.

1. *perluniprops: Index of Unicode Version 7.0.0 character properties in Perl* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/misc/perluniprops.zip)|[source](http://perldoc.perl.org/perluniprops.html)]  
   id: perluniprops; size: 100266; author: ; copyright: ; license: ;
2. *The monolingual word aligner (Sultan et al. 2015) subset of the Paraphrase Database.* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/misc/mwa_ppdb.zip)|[source](http://www.cis.upenn.edu/~ccb/ppdb/)]  
   id: mwa\_ppdb; size: 1594711; author: ; copyright: ; license: Creative Commons Attribution 3.0 Unported (CC-BY);
3. *Punkt Tokenizer Models* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/tokenizers/punkt.zip)|[source](http://www.nltk.org/nltk_data/)]  
   id: punkt; size: 13707633; author: Jan Strunk; copyright: ; license: ;
4. *RSLP Stemmer (Removedor de Sufixos da Lingua Portuguesa)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/stemmers/rslp.zip)|[source](http://www.nltk.org/nltk_data/)]  
   id: rslp; size: 3805; author: Viviane Moreira Orengo (vmorengo@inf.ufrgs.br) and Christian Huyck; copyright: ; license: ;
5. *Porter Stemmer Test Files* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/stemmers/porter_test.zip)|[source](http://www.nltk.org/nltk_data/)]  
   id: porter\_test; size: 200510; author: ; copyright: ; license: ;
6. *Snowball Data* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/stemmers/snowball_data.zip)|[source](https://github.com/snowballstem/snowball-data)]  
   id: snowball\_data; size: 6785405; author: ; copyright: ; license: ;
7. *ACE Named Entity Chunker (Maximum entropy)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/chunkers/maxent_ne_chunker.zip)|[source](http://www.nltk.org/nltk_data/)]  
   id: maxent\_ne\_chunker; size: 13404747; author: ; copyright: ; license: ;
8. *Moses Sample Models* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/models/moses_sample.zip)|[source](http://www.statmt.org/moses/?n=Moses.SampleData)]  
   id: moses\_sample; size: 10961490; author: ; copyright: ; license: ;
9. *BLLIP Parser: WSJ Model* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/models/bllip_wsj_no_aux.zip)|[source](http://nlp.stanford.edu/~mcclosky/models/)]  
   id: bllip\_wsj\_no\_aux; size: 24516205; author: ; copyright: ; license: ;
10. *Word2Vec Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/models/word2vec_sample.zip)|[source](https://code.google.com/p/word2vec/)]  
    id: word2vec\_sample; size: 49396025; author: ; copyright: ; license: ;
11. *Evaluation data from WMT15* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/models/wmt15_eval.zip)|[source](http://www.statmt.org/wmt15/)]  
    id: wmt15\_eval; size: 383096; author: ; copyright: ; license: ;
12. *Grammars for Spanish* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/grammars/spanish_grammars.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: spanish\_grammars; size: 4047; author: Kepa Sarasola; copyright: ; license: ;
13. *Sample Grammars* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/grammars/sample_grammars.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: sample\_grammars; size: 20293; author: ; copyright: ; license: ;
14. *Large context-free and feature-based grammars for parser comparison* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/grammars/large_grammars.zip)|[source](http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/elsps.html)]  
    id: large\_grammars; size: 283747; author: ; copyright: ; license: See the individual grammar files;
15. *Grammars from NLTK Book* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/grammars/book_grammars.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: book\_grammars; size: 9103; author: Ewan Klein; copyright: ; license: ;
16. *Grammars for Basque* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/grammars/basque_grammars.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: basque\_grammars; size: 4704; author: Kepa Sarasola; copyright: ; license: ;
17. *Treebank Part of Speech Tagger (Maximum entropy)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/taggers/maxent_treebank_pos_tagger.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: maxent\_treebank\_pos\_tagger; size: 10156853; author: ; copyright: ; license: ;
18. *Averaged Perceptron Tagger* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/taggers/averaged_perceptron_tagger.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: averaged\_perceptron\_tagger; size: 2526731; author: ; copyright: ; license: ;
19. *Averaged Perceptron Tagger (Russian)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/taggers/averaged_perceptron_tagger_ru.zip)|[source](http://www.ruscorpora.ru/en/)]  
    id: averaged\_perceptron\_tagger\_ru; size: 8628828; author: ; copyright: ; license: ;
20. *Mappings to the Universal Part-of-Speech Tagset* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/taggers/universal_tagset.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: universal\_tagset; size: 19095; author: ; copyright: ; license: ;
21. *VADER Sentiment Lexicon* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/sentiment/vader_lexicon.zip)|[source](https://github.com/cjhutto/vaderSentiment)]  
    id: vader\_lexicon; size: 90486; author: C.J. Hutto and Eric Gilbert; copyright: ; license: MIT License;
22. *Lin's Dependency Thesaurus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/lin_thesaurus.zip)|[source](http://webdocs.cs.ualberta.ca/~lindek/downloads.htm)]  
    id: lin\_thesaurus; size: 89154019; author: Dekang Lin; copyright: ; license: Distributed with permission of Dekang Lin;
23. *Sentiment Polarity Dataset Version 2.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/movie_reviews.zip)|[source](http://www.cs.cornell.edu/people/pabo/movie-review-data/)]  
    id: movie\_reviews; size: 4004848; author: Bo Pang and Lillian Lee; copyright: Copyright (C) 2004 Bo Pang and Lillian Lee; license: Creative Commons Attribution 4.0 International;
24. *Problem Report Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/problem_reports.zip)|[source](http://www.cs.cmu.edu/~marmalade/reports.html)]  
    id: problem\_reports; size: 1032942; author: Andrew Ko, Carnegie Mellon University; copyright: ; license: ;
25. *Pros and Cons* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/pros_cons.zip)|[source](http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets)]  
    id: pros\_cons; size: 746276; author: Bing Liu; copyright: Copyright (C) 2008 Bing Liu; license: Creative Commons Attribution 4.0 International;
26. *MASC Tagged Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/masc_tagged.zip)|[source](http://www.anc.org/)]  
    id: masc\_tagged; size: 1602143; author: Nancy Ide; copyright: Copyright (C) 2014 American National Corpus; license: This data may be used for the purposes of linguistic education, research, and development, including commercial development.;
27. *Sentence Polarity Dataset v1.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/sentence_polarity.zip)|[source](http://www.cs.cornell.edu/People/pabo/people/pabo/movie-review-data)]  
    id: sentence\_polarity; size: 490256; author: Bo Pang and Lillian Lee; copyright: Copyright (C) 2005 Bo Pang and Lillian Lee; license: Creative Commons Attribution 4.0 International;
28. *Web Text Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/webtext.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: webtext; size: 646297; author: ; copyright: ; license: ;
29. *NPS Chat* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/nps_chat.zip)|[source](http://faculty.nps.edu/cmartell/NPSChat.htm)]  
    id: nps\_chat; size: 301366; author: Craig Martell (cmartell@nps.edu); copyright: ; license: This corpus is distributed solely for non-commercial, non-profit educational and research use. It is a derivative compilation work of multiple works whose copyrights are held by the respective original authors.;
30. *City Database* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/city_database.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: city\_database; size: 1708; author: ; copyright: ; license: ;
31. *Sample European Parliament Proceedings Parallel Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/europarl_raw.zip)|[source](http://www.statmt.org/europarl)]  
    id: europarl\_raw; size: 12594977; author: Philipp Koehn, University of Edinburgh; copyright: ; license: ;
32. *BioCreAtIvE (Critical Assessment of Information Extraction Systems in Biology)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/biocreative_ppi.zip)|[source](http://www.mitre.org/public/biocreative/)]  
    id: biocreative\_ppi; size: 223566; author: ; copyright: Public Domain (not copyrighted); license: Public Domain;
33. *VerbNet Lexicon, Version 3.3* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/verbnet3.zip)|[source](https://verbs.colorado.edu/verbnet/)]  
    id: verbnet3; size: 482025; author: Karin Kipper-Schuler; copyright: ; license: Distributed with permission of the author.;
34. *Cross-Framework and Cross-Domain Parser Evaluation Shared Task* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/pe08.zip)|[source](http://www-tsujii.is.s.u-tokyo.ac.jp/pe08-st/)]  
    id: pe08; size: 80735; author: ; copyright: ; license: Distributed with permission;
35. *The Patient Information Leaflet (PIL) Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/pil.zip)|[source](http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/)]  
    id: pil; size: 1510205; author: ; copyright: ; license: Distributed with permission;
36. *Crubadan Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/crubadan.zip)|[source](http://borel.slu.edu/crubadan/)]  
    id: crubadan; size: 5288655; author: Kevin Scannell; copyright: Copyright (C) 2010 Kevin Scannell; license: GPLv3;
37. *Project Gutenberg Selections* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/gutenberg.zip)|[source](http://gutenberg.net/)]  
    id: gutenberg; size: 4251829; author: ; copyright: public domain; license: public domain;
38. *Proposition Bank Corpus 1.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/propbank.zip)|[source](http://verbs.colorado.edu/~mpalmer/projects/ace.html)]  
    id: propbank; size: 5323498; author: ; copyright: ; license: Distributed with permission;
39. *Machado de Assis -- Obra Completa* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/machado.zip)|[source](http://machado.mec.gov.br/)]  
    id: machado; size: 6151774; author: Machado de Assis; copyright: ; license: Public Domain;
40. *C-Span State of the Union Address Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/state_union.zip)|[source](http://www.c-span.org/executive/stateoftheunion.asp)]  
    id: state\_union; size: 808757; author: ; copyright: public domain; license: public domain;
41. *Twitter Samples* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/twitter_samples.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: twitter\_samples; size: 16007673; author: ; copyright: Copyright (C) 2015 Twitter, Inc; license: Must be used subject to Twitter Developer Agreement (https://dev.twitter.com/overview/terms/agreement);
42. *SemCor 3.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/semcor.zip)|[source](http://www.cse.unt.edu/~rada/downloads.html#semcor)]  
    id: semcor; size: 4397021; author: Rada Mihalcea (rada@cs.unt.edu); copyright: ; license: You are granted permission to use, copy, modify and distribute this database for any purpose and without fee and royalty is hereby granted, provided that you agree to comply with the Princeton copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the database, including modifications that you make for internal use or for distribution. See semcor/README for more information.;
43. *Names Corpus, Version 1.3 (1994-03-29)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/names.zip)|[source](http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/)]  
    id: names; size: 21326; author: Mark Kantrowitz and Bill Ross; copyright: Copyright (C) 1991 Mark Kantrowitz; license: You may use the lists of names for any purpose, so long as credit is given in any published work. You may also redistribute the list if you provide the recipients with a copy of this README file. The lists are not in the public domain (I retain the copyright on the lists) but are freely redistributable. If you have any additions to the lists of names, I would appreciate receiving them.;
44. *Penn Treebank* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/ptb.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: ptb; size: 6289; author: ; copyright: Copyright (C) 1995 University of Pennsylvania; license: This is a stub for the full Penn Treebank Corpus version 3.;
45. *NomBank Corpus 1.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/nombank.1.0.zip)|[source](http://nlp.cs.nyu.edu/meyers/NomBank.html)]  
    id: nombank.1.0; size: 6728397; author: ; copyright: ; license: Distributed with permission;
46. *Portuguese Treebank* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/floresta.zip)|[source](http://www.linguateca.pt/Floresta/)]  
    id: floresta; size: 1882021; author: ; copyright: ; license: Non-commercial use only;
47. *ComTrans Corpus Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/comtrans.zip)|[source](http://www.fask.uni-mainz.de/user/rapp/comtrans/)]  
    id: comtrans; size: 11904518; author: Reinhard Rapp; copyright: ; license: ;
48. *KNB Corpus (Annotated blog corpus)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/knbc.zip)|[source](http://lilyx.net/pages/nltkjapanesecorpus.html)]  
    id: knbc; size: 8760788; author: ; copyright: ; license: Freely re-distributable under the same license as the original KNB Corpus.;
49. *MAC-MORPHO: Brazilian Portuguese news text with part-of-speech tags* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/mac_morpho.zip)|[source](http://www.nilc.icmc.usp.br/lacioweb/)]  
    id: mac\_morpho; size: 3013904; author: ; copyright: ; license: Distributed with permission of Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade de São Paulo (USP) in São Carlos, Universidade Federal de São Carlos (UFSCar), Universidade Estadual Paulista (UNESP) of Araraquara.;
50. *Swadesh Wordlists* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/swadesh.zip)|[source](http://en.wiktionary.org/wiki/Appendix:Swadesh_list)]  
    id: swadesh; size: 22828; author: ; copyright: ; license: GNU Free Documentation License;
51. *PASCAL RTE Challenges 1, 2, and 3* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/rte.zip)|[source](http://www.pascal-network.org/Challenges/RTE/)]  
    id: rte; size: 386303; author: ; copyright: ; license: ;
52. *Toolbox Sample Files* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/toolbox.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: toolbox; size: 250616; author: ; copyright: ; license: ;
53. *JEITA Public Morphologically Tagged Corpus (in ChaSen format)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/jeita.zip)|[source](http://lilyx.net/pages/nltkjapanesecorpus.html)]  
    id: jeita; size: 16531215; author: ; copyright: ; license: Freely re-distributable under the same license as the original JEITA corpus. Each document retains its own license from Aozora bunko and Project Sugita Genpaku.;
54. *Product Reviews (5 Products)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/product_reviews_1.zip)|[source](http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets)]  
    id: product\_reviews\_1; size: 141287; author: Bing Liu; copyright: Copyright (C) 2004 Bing Liu; license: Creative Commons Attribution 4.0 International;
55. *Open Multilingual Wordnet* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/omw.zip)|[source](http://compling.hss.ntu.edu.sg/omw/)]  
    id: omw; size: 12110409; author: Francis Bond; copyright: Please consult the copyright statements of the individual Wordnets; license: Please consult the LICENSE files included with the individual Wordnets. Note that all permit redistribution.;
56. *SentiWordNet* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/sentiwordnet.zip)|[source](http://sentiwordnet.isti.cnr.it/)]  
    id: sentiwordnet; size: 4686546; author: Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani; copyright: Copyright (C) 2013 SentiWordNet Project; license: Creative Commons Attribution ShareAlike 3.0 Unported license;
57. *Product Reviews (9 Products)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/product_reviews_2.zip)|[source](http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets)]  
    id: product\_reviews\_2; size: 170698; author: Bing Liu; copyright: Copyright (C) 2007 Bing Liu; license: Creative Commons Attribution 4.0 International;
58. *Australian Broadcasting Commission 2006* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/abc.zip)|[source](http://www.abc.net.au/)]  
    id: abc; size: 1487851; author: Australian Broadcasting Commission; copyright: ; license: ;
59. *Universal Declaration of Human Rights Corpus (Unicode Version)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/udhr2.zip)|[source](http://unicode.org/udhr/)]  
    id: udhr2; size: 1653975; author: ; copyright: public domain; license: public domain;
60. *SENSEVAL 2 Corpus: Sense Tagged Text* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/senseval.zip)|[source](http://www.senseval.org/)]  
    id: senseval; size: 2151350; author: ; copyright: ; license: Distributed with permission.;
61. *Word Lists* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/words.zip)|[source](http://en.wikipedia.org/wiki/Words_(Unix))]  
    id: words; size: 757777; author: ; copyright: public domain; license: public domain;
62. *FrameNet 1.5* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/framenet_v15.zip)|[source](http://framenet.icsi.berkeley.edu/)]  
    id: framenet\_v15; size: 69337891; author: Collin F. Baker; copyright: ; license: May be used for non-commercial purposes.;
63. *Unicode Samples* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/unicode_samples.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: unicode\_samples; size: 1212; author: ; copyright: ; license: ;
64. *PC-KIMMO Data Files* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/kimmo.zip)|[source](http://www.sil.org/pckimmo/)]  
    id: kimmo; size: 186958; author: ; copyright: ; license: ;
65. *FrameNet 1.7* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/framenet_v17.zip)|[source](http://framenet.icsi.berkeley.edu/)]  
    id: framenet\_v17; size: 99207152; author: Collin F. Baker; copyright: ; license: Creative Commons Attribution 3.0 Unported License;
66. *Chat-80 Data Files* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/chat80.zip)|[source](http://www.cis.upenn.edu/~pereira/oldies.html)]  
    id: chat80; size: 19209; author: David Warren and Fernando Pereira; copyright: Copyright (C) 1982 David Warren and Fernando Pereira; license: This program may be used, copied, altered or included in other programs only for academic purposes and provided that the authorship of the initial program is aknowledged. Use for commercial purposes without the previous written agreement of the authors is forbidden.;
67. *Experimental Data for Question Classification* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/qc.zip)|[source](http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/)]  
    id: qc; size: 125456; author: Xin Li and Dan Roth, UIUC; copyright: ; license: ;
68. *C-Span Inaugural Address Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/inaugural.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: inaugural; size: 329806; author: ; copyright: public domain; license: public domain;
69. *WordNet* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/wordnet.zip)|[source](http://wordnet.princeton.edu/)]  
    id: wordnet; size: 10775600; author: ; copyright: WordNet 3.0 Copyright 2006 by Princeton University. All rights reserved.; license: Permission to use, copy, modify and distribute this software and database and its documentation for any purpose and without fee or royalty is hereby granted, provided that you agree to comply with the following copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the software, database and documentation, including modifications that you make for internal use or for distribution.... [see webpage for full license];
70. *Stopwords Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/stopwords.zip)|[source](ftp://ftp.cs.cornell.edu/pub/smart/english.stop%20and%20http:/snowball.tartarus.org/%20and%20others)]  
    id: stopwords; size: 23047; author: ; copyright: ; license: ;
71. *VerbNet Lexicon, Version 2.1* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/verbnet.zip)|[source](https://verbs.colorado.edu/verbnet/)]  
    id: verbnet; size: 323661; author: Karin Kipper-Schuler; copyright: ; license: Distributed with permission of the author.;
72. *Shakespeare XML Corpus Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/shakespeare.zip)|[source](http://www.andrew.cmu.edu/user/akj/shakespeare/)]  
    id: shakespeare; size: 475458; author: ; copyright: public domain; license: public domain;
73. *York-Toronto-Helsinki Parsed Corpus of Old English Prose* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/ycoe.zip)|[source](http://www.ota.ahds.ac.uk/)]  
    id: ycoe; size: 477; author: ; copyright: ; license: ;
74. *NIST IE-ER DATA SAMPLE* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/ieer.zip)|[source](http://www.itl.nist.gov/iad/894.01/tests/ie-er/er_99/er_99.htm)]  
    id: ieer; size: 166156; author: ; copyright: ; license: ;
75. *CESS-CAT Treebank* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/cess_cat.zip)|[source](http://clic.ub.edu/cessece/)]  
    id: cess\_cat; size: 5396688; author: ; copyright: ; license: If you use these corpora for research, please cite thusly: CESS-Cat project (M. Antonia Martí, MarionaTaulé, Lluís Márquez, Manuel Bertran (2007) ?CESS-ECE: A Multilingual and Multilevel Annotated Corpus? in http://www.lsi.upc.edu/~mbertran/cess-ece/publications).;
76. *Switchboard Corpus Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/switchboard.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: switchboard; size: 791161; author: ; copyright: ; license: Permission is granted for use of this material in accordance with the Open Content License [http://opencontent.org/opl.shtml]. This corpus contains transcripts and annotations for 36 calls from the Switchboard Corpus [http://www.ldc.upenn.edu/Catalog/LDC93S7.html].;
77. *Comparative Sentence Dataset* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/comparative_sentences.zip)|[source](http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets)]  
    id: comparative\_sentences; size: 279121; author: Nitin Jindal and Bing Liu; copyright: Copyright (C) 2006 Nitin Jindal and Bing Liu; license: Creative Commons Attribution 4.0 International;
78. *Subjectivity Dataset v1.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/subjectivity.zip)|[source](http://www.cs.cornell.edu/People/pabo/people/pabo/movie-review-data)]  
    id: subjectivity; size: 521628; author: Bo Pang and Lillian Lee; copyright: Copyright (C) 2004 Bo Pang and Lillian Lee; license: Creative Commons Attribution 4.0 International;
79. *Universal Declaration of Human Rights Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/udhr.zip)|[source](http://www.un.org/Overview/rights.html)]  
    id: udhr; size: 1170177; author: ; copyright: public domain; license: public domain;
80. *Polish language of the XX century sixties* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/pl196x.zip)|[source](http://www.mimuw.edu.pl/polszczyzna/pl196x/index_en.htm)]  
    id: pl196x; size: 7051453; author: I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak; copyright: ; license: GNU General Public License;
81. *Paradigm Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/paradigms.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: paradigms; size: 24902; author: Cathy Bow, University of Melbourne; copyright: ; license: Distributed with the permission of the author;
82. *Gazeteer Lists* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/gazetteers.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: gazetteers; size: 8265; author: ; copyright: ; license: GNU Free Documentation License; or public domain (depending on the file);
83. *TIMIT Corpus Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/timit.zip)|[source](http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1)]  
    id: timit; size: 22251869; author: ; copyright: ; license: This corpus sample is Copyright 1993 Linguistic Data Consortium, and is distributed under the terms of the Creative Commons Attribution, Non-Commercial, ShareAlike license. http://creativecommons.org/;
84. *Penn Treebank Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/treebank.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: treebank; size: 1740034; author: ; copyright: Copyright (C) 1995 University of Pennsylvania; license: This is a 10% fragment of Penn Treebank, (C) LDC 1995. It is made available under fair use for the purposes of illustrating NLTK tools for tokenizing, tagging, chunking and parsing. This data is for non-commercial use only.;
85. *Sinica Treebank Corpus Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/sinica_treebank.zip)|[source](http://rocling.iis.sinica.edu.tw/CKIP/engversion/treebank.htm)]  
    id: sinica\_treebank; size: 899237; author: ; copyright: ; license: Distributed with the Natural Language Toolkit under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike License [http://creativecommons.org/licenses/by-nc-sa/2.5/].;
86. *Opinion Lexicon* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/opinion_lexicon.zip)|[source](http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets)]  
    id: opinion\_lexicon; size: 24947; author: Bing Liu; copyright: Copyright (C) 2011 Bing Liu; license: Creative Commons Attribution 4.0 International;
87. *Prepositional Phrase Attachment Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/ppattach.zip)|[source](ftp://ftp.cis.upenn.edu/pub/adwait/PPattachData/)]  
    id: ppattach; size: 781714; author: Adwait Ratnaparkhi; copyright: (C) 1994 Adwait Ratnaparkhi; license: Distributed with the permission of the author.;
88. *Dependency Parsed Treebank* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/dependency_treebank.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: dependency\_treebank; size: 457429; author: ; copyright: Copyright (C) 1995 University of Pennsylvania; license: This is a 10% fragment of Penn Treebank, (C) LDC 1995, which has been dependency parsed. It is made available under fair use for the purposes of illustrating NLTK tools for tokenizing, tagging, chunking and parsing. This data is for non-commercial use only.;
89. *The Reuters-21578 benchmark corpus, ApteMod version* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/reuters.zip)|[source](http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html)]  
    id: reuters; size: 6378691; author: ; copyright: ; license: The copyright for the text of newswire articles and Reuters annotations in the Reuters-21578 collection resides with Reuters Ltd. Reuters Ltd. and Carnegie Group, Inc. have agreed to allow the free distribution of this data \*for research purposes only\*. If you publish results based on this data set, please acknowledge its use, refer to the data set by the name 'Reuters-21578, Distribution 1.0', and inform your readers of the current location of the data set.;
90. *Genesis Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/genesis.zip)|[source](http://www.nltk.org/nltk_data/)]  
    id: genesis; size: 473239; author: ; copyright: public domain; license: public domain;
91. *CESS-ESP Treebank* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/cess_esp.zip)|[source](http://clic.ub.edu/cessece/)]  
    id: cess\_esp; size: 2220392; author: ; copyright: ; license: If you use these corpora for research, please cite thusly: CESS-Cat project (M. Antonia Martí, MarionaTaulé, Lluís Márquez, Manuel Bertran (2007) ?CESS-ECE: A Multilingual and Multilevel Annotated Corpus? in http://www.lsi.upc.edu/~mbertran/cess-ece/publications).;
92. *Dependency Treebanks from CoNLL 2007 (Catalan and Basque Subset)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/conll2007.zip)|[source](http://nextens.uvt.nl/depparse-wiki/DataDownload)]  
    id: conll2007; size: 1242958; author: ; copyright: Copyright (C) 2007 The University of the Basque Country; license: Creative Commons Attribution-NonCommercial-NoDerivativeWorks license;
93. *Non-Breaking Prefixes (Moses Decoder)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/nonbreaking_prefixes.zip)|[source](https://github.com/moses-smt/mosesdecoder/tree/master/scripts/share/nonbreaking_prefixes)]  
    id: nonbreaking\_prefixes; size: 25437; author: ; copyright: ; license: Gnu LGPL;
94. *Dolch Word List* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/dolch.zip)|[source](https://en.wikipedia.org/wiki/Dolch_word_list)]  
    id: dolch; size: 2116; author: ; copyright: ; license: ;
95. *SMULTRON Corpus Sample* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/smultron.zip)|[source](http://www.ling.su.se/DaLi/research/smultron/index.htm)]  
    id: smultron; size: 166207; author: Sofia Gustafson-Capkova, Yvonne Samuelsson, and Martin Volk; copyright: ; license: ;
96. *Alpino Dutch Treebank* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/alpino.zip)|[source](http://www.let.rug.nl/~vannoord/trees/)]  
    id: alpino; size: 2797255; author: ; copyright: ; license: Distributed with permission of Gertjan van Noord;
97. *WordNet-InfoContent* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/wordnet_ic.zip)|[source](http://wn-similarity.sourceforge.net/)]  
    id: wordnet\_ic; size: 12056682; author: ; copyright: ; license: ;
98. *Brown Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/brown.zip)|[source](http://www.hit.uib.no/icame/brown/bcm.html)]  
    id: brown; size: 3314357; author: W. N. Francis and H. Kucera; copyright: ; license: May be used for non-commercial purposes.;
99. *PanLex Swadesh Corpora* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/panlex_swadesh.zip)|[source](http://panlex.org/)]  
    id: panlex\_swadesh; size: 2861668; author: Jonathan Pool (editor); copyright: ; license: CC0 1.0 Universal;
100. *CONLL 2000 Chunking Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/conll2000.zip)|[source](http://www.cnts.ua.ac.be/conll2000/chunking/)]  
     id: conll2000; size: 756607; author: ; copyright: ; license: ;
101. *Universal Treebanks Version 2.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/universal_treebanks_v20.zip)|[source](https://code.google.com/p/uni-dep-tb/)]  
     id: universal\_treebanks\_v20; size: 25908853; author: ; copyright: ; license: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States;
102. *Brown Corpus (TEI XML Version)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/brown_tei.zip)|[source](http://www.hit.uib.no/icame/brown/bcm.html)]  
     id: brown\_tei; size: 8737738; author: W. N. Francis and H. Kucera; copyright: ; license: May be used for non-commercial purposes.;
103. *The Carnegie Mellon Pronouncing Dictionary (0.6)* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/cmudict.zip)|[source](ftp://ftp.cs.cmu.edu/project/speech/dict/)]  
     id: cmudict; size: 896069; author: ; copyright: Copyright 1998 Carnegie Mellon University; license: Use of this dictionary, for any research or commercial purpose, is completely unrestricted. If you use or redistribute this material, we would appreciate acknowlegement of its origin.;
104. *MULTEXT-East 1984 annotated corpus 4.0* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/mte_teip5.zip)|[source](https://www.clarin.si/repository/xmlui/handle/11356/1043)]  
     id: mte\_teip5; size: 14800561; author: Erjavec, Tomaž; Barbu, Ana-Maria; Derzhanski, Ivan; Dimitrova, Ludmila; Garabík, Radovan; Ide, Nancy; Kaalep, Heiki-Jaan; Kotsyba, Natalia; Krstev, Cvetana; Oravecz, Csaba; Petkevič, Vladimír; Priest-Dorman, Greg; QasemiZadeh, Behrang; Radziszewski, Adam; Simov, Kiril; Tufiş, Dan and Zdravkova, Katerina; copyright: ; license: Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0);
105. *Indian Language POS-Tagged Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/indian.zip)|[source](http://www.nltk.org/nltk_data/)]  
     id: indian; size: 199187; author: A Kumaran; copyright: ; license: Distributed with permission;
106. *CONLL 2002 Named Entity Recognition Corpus* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/conll2002.zip)|[source](http://www.cnts.ua.ac.be/conll2002/ner/)]  
     id: conll2002; size: 1867449; author: ; copyright: ; license: ;
107. *Help on Tagsets* [[download](https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/help/tagsets.zip)|[source](http://www.nltk.org/nltk_data/)]  
     id: tagsets; size: 34531; author: UCREL, Lancaster University; copyright: ; license: ;