MANUAL OF INFORMATION

to accompany

A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.

by

W. N. Francis

H. Kučera

Brown University

Providence, Rhode Island
Department of Linguistics
Brown University
1964

Revised 1971

Revised and Amplified
1979

MANUAL OF INFORMATION

to accompany

A Standard Corpus of Present-Day

Edited American English, for use

with Digital Computers.

by

W. N. Francis

H. Kučera

Brown University

Providence, Rhode Island

Department of Linguistics

Brown University

1964

Revised 1971

Revised and Amplified

1979

PREFACE

To Revised Edition, 1979

This Manual was first published in 1964, when the Standard Sample of Present-Day American English (the Brown Corpus) was first made available.* A revised edition was issued in 1971, principally to incorporate information about the text turned up in seven years of use. The present revision is more extensive, since it includes information about recently prepared versions of the Corpus, notably the "tagged" text completed at Brown University in 1979. Two complete proofreadings of the Corpus have resulted in corrections of two kinds: errors in the preparation of the original tape, which have been silently corrected in recently issued copies, and further typographical errors and anomalies in the underlying text, which have been recorded in the descriptions of individual samples on pages 33-176. (Most of these were listed on corrigenda sheets which have been enclosed with recently issued copies of the Manual.)

We wish to record here our thanks to all those who have sent in information about errors in the Corpus, and our special gratitude to those who have worked on the production of alternate versions of the Corpus, notably Gerald M. Rubin, Barbara Greene Levine, Sandra Pearce, Patricia Strauss, Stephen Ritz, Andrew Mackie, Jostein Hauge, and Donald Sherman (a partial list). At the time of writing, more than 160 copies

---

*The original Standard Corpus was prepared under a grant from the Cooperative Research Program of the U.S. Office of Education.

of the Corpus are in circulation, and a recent bibliography of published works using or referring to the Corpus includes 57 items (ICAME News, No. 2, Bergen, March 1979, pp. 9-12).

W. Nelson Francis

Henry Kučera

Brown University

July 1979.

TABLE OF CONTENTS

1. CONTENTS

This Standard Corpus of Present-Day American English consists of 1,014,312 words[1] of running text of edited English prose printed in the United States during the calendar year 1961. So far as it has been possible to determine, the writers were native speakers of American English. Although all of the material first appeared in print in the year 1961, some of it was undoubtedly written earlier. However, no material known to be a second edition or reprint of earlier text has been included.

The Corpus is divided into 500 samples of 2000+ words each. Each sample begins at the beginning of a sentence but not necessarily of a paragraph or other larger division, and each ends at the first sentence-ending after 2000 words.[2] The samples represent a wide range of styles and varieties of prose. Verse was not included on the ground that it presents special linguistic problems different from those of prose. (Short verse passages quoted in prose samples are kept, however.) Drama was excluded as being the imaginative recreation of spoken discourse, rather than true written discourse. Fiction was included, but no samples were admitted which consisted of more than 50% dialogue. Samples were chosen for their representative quality rather than for any subjectively determined excellence. The use of the word *standard* in the title of the Corpus does not in any way mean that it is put forward as "standard

---

[1] Owing to subsequent corrections, this total may be in error by as much as 20, plus or minus.

[2] In a few cases the count erroneously extended over this limit, but the extra material has been allowed to remain. Owing to errors in the rough count, 15 samples have between 1,990 and 1,999 words, and 3 have fewer than 1,990. The average length is 2,028.6.

English"; it merely expresses the hope that this corpus will be used for comparative studies where it is important to use the same body of data. Since the preparation and input of data is a major bottleneck in computer work, the intent was to make available a carefully chosen and prepared body of material of considerable size in standardized format. The corpus may further prove to be standard in setting the pattern for the preparation and presentation of further bodies of data in English or in other languages.[3]

The selection procedure was in two phases: an initial subjective classification and decision as to how many samples of each category would be used, followed by a random selection of the actual samples within each category. In most categories the holdings of the Brown University Library and the Providence Athenaeum were treated as the universe from which the random selections were made. But for certain categories it was necessary to go beyond these two collections. For the daily press, for example, the list of American newspapers of which the New York Public Library keeps microfilm files was used (with the addition of the *Providence Journal*). Certain categories of chiefly ephemeral material necessitated rather arbitrary decisions; some periodical materials in the categories *Skills and Hobbies* and *Popular Lore* were chosen from the contents of one of the largest second-hand magazine stores in New York City.

The list of main categories and their subdivisions was drawn up at a conference held at Brown University in February 1963.[4] The participants

---

[3]This expectation has been realized by the preparation of a corpus of British English replicating as closely as possible the format of the Brown Corpus—the Lancaster-Oslo-Bergen Corpus, prepared by Geoffrey Leech and Stig Johansson.

[4]Participants in the conference were John B. Carroll, W. Nelson Francis, Philip B. Gove, Henry Kučera, Patricia O'Connor, and Randolph Quirk.

in the conference also independently gave their opinions as to the number of samples there should be in each category. These figures were averaged to obtain the preliminary set of figures used. A few changes were later made on the basis of experience gained in making the selections. Finer subdivision was based on proportional amounts of actual publication during 1961.[5] The list of main categories with their principal subdivisions and the number of samples in each follows:

I. Informative Prose — 374 samples

A. Press: Reportage

	Daily	Weekly	Total	
Political	10	4	14	
Sports	5	2	7	
Society	3	0	3	
Spot News	7	2	9	
Financial	3	1	4	
Cultural	5	2	7	
Total				44

B. Press: Editorial

	Daily	Weekly	Total	
Institutional	7	3	10	
Personal	7	3	10	
Letters to the Editor	5	2	7	
Total				27

C. Press: Reviews (theatre, books, music, dance)

Daily	Weekly	Total	
14	3		17

D. Religion

Books	7
Periodicals	6
Tracts	4
Total	17

E. Skills and Hobbies

Books	2
Periodicals	34
Total	36

F. Popular Lore

Books	23
Periodicals	25
Total	48

[5]Determined where possible from the monthly *American Book Publishing Record*, published by R. R. Bowker Company, New York, New York.

I. Informative Prose, continued

G. Belles Lettres, Biography, Memoirs, etc.

Books	38	
Periodicals	37	
Total		75

H. Miscellaneous

Government Documents	24	
Foundation Reports	2	
Industry Reports	2	
College Catalog	1	
Industry House Organ	1	
Total		30

J. Learned

Natural Sciences	12	
Medicine	5	
Mathematics	4	
Social and Behavioral Sciences	14	
Political Science, Law, Education	15	
Humanities	18	
Technology and Engineering	12	
Total		80

II. Imaginative Prose 126 Samples

K. General Fiction

Novels	20	
Short Stories	9	
Total		29

L. Mystery and Detective Fiction

Novels	20	
Short Stories	4	
Total		24

M. Science Fiction

Novels	3	
Short Stories	3	
Total		6

N. Adventure and Western Fiction

Novels	15	
Short Stories	14	
Total		29

P. Romance and Love Story

Novels	14	
Short Stories	15	
Total		29

II. Imaginative Prose, continued

R. Humor

Novels	3
Essays, etc.	6
Total	9
GRAND TOTAL	500

Once these categories, subcategories, and numbers of samples had been decided upon, the choice of the actual samples was made by various random methods, chiefly the use of a table of random numbers applied to the total list of available publications in the subject field in question. The page on which to begin the sample was also selected by the random number table. Each sample begins with the first complete sentence on the page so selected. Titles and running heads have been omitted, also footnotes, tables, and picture captions. A rough count of 2,000 words was made and the sample was terminated at the next sentence-break. For purposes of this count a word was defined as any string of characters which would appear in the final coding (to be described below) with space on either side, except for the code indicators of paragraph beginnings. A sentence was defined as a string of such words beginning with a capital and ending with a final mark (. ! or ?) followed by space and a capital, excluding obvious abbreviations. In some cases the sentence final mark may not be followed by space; see under *Quotations* below. Subsequently a more accurate count was made by the computer.

Each sample was given a code number consisting of the letter designating its category on the list above followed by a two-digit number. Thus A01 is the first sample under *Press: Reportage* and G75 is the last under *Belles Lettres*. No special order was followed within each category, though in most cases the members of each subcategory appear in

sequence. In general, however, the sequence within each major category was determined by the order of selection.

For all copyrighted material used, the permission of the copyright holder has been obtained. Details of copyright permission are included in the separate listing of the samples on pages 33-176.

2. VERSIONS OF THE CORPUS

Six versions of the Corpus are available. All contain the same basic text, but they differ in typography and format.

(1) Form A. This is the original form of the Corpus, as it was prepared in 1963-64. The limitations of computer printing facilities at that time required that it use an elaborate coding procedure, which is described in Section 3 below.

(2) Form B. This is the "stripped" version, from which all punctuation symbols and codes except hyphens, apostrophes, and symbols for formulas and ellipses have been omitted. It is especially useful for those who are interested in individual words, and was used in the preparation of the frequency tables in Kučera and Francis, *Computational Analysis of Present-Day American English* (Providence: Brown University Press, 1967).

(3) Form C, This is the "tagged" version, which makes use of a partially stripped text in which only proper name capitalization and those punctuation marks which are of grammatical significance have been retained. Each individual word (token) in this version has been given a grammatical tag from a list of 81, each specifying a particular word-class. The nature and rationale of this tagging are explained in full in Section 4 below.

(4) Bergen Form I. This version and the following were prepared at the Norwegian Computational Center for Humanistic Research (NAVF's EDB-senter for humanistisk forskning) at the University of Bergen under the direction of Dr. Jostein Hauge. Both contain upper- and lower-case letters, regular punctuation marks, and a minimum of special codes. In this version, typographic information is preserved and the same line division is used as in the original version except that words at the end of the line are never divided.

(5) Bergen Form II. In this version typographical information is somewhat reduced and a new longer line is used. This version is available on microfiche, together with a complete KWIC concordance, from the EDB-senter (P.O. Box 53, University of Bergen, N-5014 Bergen, Norway).

(6) Brown MARC Form. This version was prepared at Stanford University. It is designed to be compatible with two commonly used research techniques which are appropriate for large textual corpora:

(1) searching for and retrieving full-sentence citations using single words or word + context as retrieval criteria;

(2) generating KWIC-form concordances which can be organized according to varying arrangements of a keyword plus its preceding or following verbal context.

This is thus a variable-length record format, using the sentence as a single record. It is available from the Stanford Computer Archive of Language Materials (CALM), Department of Linguistics, Stanford University, Stanford, California 94305.

3. CODING PROCEDURE OF FORM A

The basic coding procedure followed in Form A is that devised for

the U. S. Patent Office and described in the pamphlet *A Notation System for Transliterating Technical and Scientific Texts for Use in Data Processing Systems*, by Simon M. Newman, Rowena W. Swanson, and Kenneth Knowlton (Patent Office Research and Development Reports, No. 15, U. S. Department of Commerce, 1959). This was the only complete coding system that we could find in 1963. But since this system was devised for the relatively stereotyped format of U. S. patents, and since it was our intent to include as much as possible of the graphic detail of the basic texts, it was necessary to make several modifications in the system. Accordingly, the whole system, as modified, will be briefly described here; users of the Corpus may refer to the Patent Office publication for full details.

The basic unit of text is the single line of 80 spaces, contained on a single IBM punched card, transferred in card image to the magnetic tape. The first 70 spaces contain text and the last nine a location marker uniquely identifying that line of text. Space 71 is blank.

Location Marker. The first four spaces of the location marker contain a line number, treated as a three-digit number with one decimal position. Normally the fourth digit is 0, but if it was necessary to correct punching errors by inserting one or more cards into the text after the original numbering was made, the fourth digit was used.[6] The next two spaces of all location markers contain the characters E1, the label of this corpus (English Corpus No. 1). The last three spaces of the location marker contain the code designation of the individual sample. A new set of card numbers is begun for each sample. Thus 0010E1A01

---

[6]Very rarely are there two or more correction cards in sequence. If so, their decimal numbers are in sequence. Single correction cards sometimes have the decimal 1 and sometimes 5.

designates the first line of sample A01, hence the first line of the Corpus. 0185E1G75 designates a line inserted between 0180 and 0190 of sample G75. The last line of each sample is blank.

Text. The text normally occupies spaces 1-70 of each line, running continuously from line to line without regard to word endings. No hyphens are used when a word is broken at the end of a line; it is simply run over without space onto the next line. If a word ends at space 70 of a line, the first space of the next line is left blank. If a word ends at space 69, space 70 is left blank and the next line normally begins on space 1. However, owing to corrections, many lines are not filled out. In this case two conventions have been used: (1) any number of blank spaces from 1 to 68 is considered a single space, and (2) the character * followed by space means "cancel the * and all following spaces." Thus text reading

```
TELEVISION IMPULSES, SOUND WAVES, ULTRA-VIOLET RAYS, ETC**., THAT MAY  1020E1F03
OC*                                                                    1025E1F03
CUPY THE VERY SAME SPACE, EACH SOLITARY UPON ITS OWN FREQUENCY, IS INF 1030E1F03
INITE. *SO WE MAY CONCEIVE THE COEXISTENCE OF THE INFINITE NUMBER OF U 1040E1F03
NIVERSAL, APPARENTLY MOMENTARY  STATES OF MATTER, SUCCESSIVE ONE AFTER 1050E1F03
 ANOTHER IN CONSCIOUSNESS, BUT PERMANENT EACH ON ITS OWN BASIC PHASE O 1060E1F03
F THE PROGRESSIVE FREQUENCIES. *THIS THEORY MAKES IT POSSIBLE FOR ANY  1070E1F03
EVENT THROUGHOUT ETERNITY TO BE CONTINUOUSLY AVAILABLE AT ANY MOMENT T 1080E1F03
O CONSCIOUSNESS. *                                                     1090E1F03
```

is to be read as transliterating

> television impulses, sound waves, ultra-violet rays, etc., that may occupy the very same space, each solitary upon its own frequency, is infinite. So we may conceive the coexistence of the infinite number of universal, apparently momentary states of matter, successive one after another in consciousness, but permanent each on its own basic phase of the progressive frequencies. This theory makes it possible for any event throughout eternity to be continuously available at any moment to consciousness.

Headings and Paragraph Divisions. The coding **N is used to mean "begin major heading" and the coding **P to mean "end major heading." By "major heading" is meant the heading of the largest subdivision of the text that falls within the sample. If a whole sample falls within a single chapter of a book, for example, these codings are used for the largest subheads, if any, within the chapter. But if a sample straddles a chapter break, these codings are used for the chapter heading. Where there is a major break in the text without a heading, these symbols are run together as **N**P.

Minor headings are indicated by the beginning symbol **R and the closing symbol **T. Even in the case of elaborately subdivided material, these symbols are used for all subdivisions below the largest, down to and including the paragraph. The beginning of a minor subdivision without heading is indicated by **R**T.

No indications of capitals, italics, and other graphic features are made within headings.

Special Types. A passage in italics is marked by the beginning symbol *= and the closing symbol *$. If the italicized passage is smaller than a word, these symbols are included in the word without spacing; thus in*cred*ible would be coded IN*=CRED*$IBLE. Similarly, bold-faced type is indicated by the begin and close symbols **= and **$. Initial capitals are indicated by a prefixed * (Since this is not followed by a blank, it will not cancel itself.) Passages of more than a single letter in capitals, whether large or small, are marked by the begin symbol **(and the close symbol **) .

Greek letters are marked by a preceding **Y for lower case and **Z

for upper case, according to the following table of equivalents

alpha	A	α	A
beta	B	β	B
gamma	Γ	γ	G
delta	Δ	δ	D
epsilon	E	ε	E
zeta	Z	ζ	Z
eta	H	η	H
theta	Θ	θ	J
iota	I	ι	I
kappa	K	κ	K
lambda	Λ	λ	L
mu	M	μ	M
nu	N	ν	N
xi	Ξ	ξ	X
omicron	O	ο	O
pi	Π	π	P
rho	P	ρ	R
sigma	Σ	σ	S
tau	T	τ	T
upsilon	Y	υ	U
phi	Φ	φ	F
chi	X	χ	C
psi	Ψ	ψ	Y
omega	Ω	ω	Q

Characters in other alphabets, type ornaments, and other typographical features for which no coding is supplied are indicated by the cover symbol **B .

Abbreviations. The period marking an abbreviation is coded **. to distinguish it from the ordinary sentence-ending period. When an abbreviation comes at the end of a sentence, the coding **.. is used to indicate the double function of the period. Abbreviations not marked by a period are treated as *symbols* (see below).

Symbols. Combinations of letters without a following period (except at the end of a sentence) and not constituting a genuine word are defined as symbols and are preceded by the code marker **J. The domain of this marker continues to the next space or mark of punctuation (including hyphen).

Formulas. Combinations of letters, numbers, and other symbols which also include operator symbols (such as +, =, exponents, subscripts) are defined as formulas and are replaced by the code **F. A formula, no matter how long, is thus counted as a single word in the word count.

Numbers. Numbers which are not part of formulas are reproduced normally, including the decimal point, which is distinguished from the period by the fact that it is not followed by space. The superscript ^{o} for degrees of angle or temperature is indicated by *+0. Superscript numbers, letters, or other characters indicating footnotes are ignored. All other superscripts and all subscripts are within formulas and thus are omitted.

Roman numerals are indicated by the begin and close symbols */ and *, with the equivalent Arabic numeral placed between them. Thus */8*, stands for VIII.

Punctuation. The following codes are used for punctuation and other graphic features:

**A = ' (apostrophe, single quote)	**U = " (end quotation)
**B = uncoded character(s)	**X = !
**C = :	+ = + or &
**D = ¨ (diaeresis or umlaut)	*(= [
**F = formula	*) =]
**I = ?	*$ = end italics or underscoring

**J = begin symbol

**K = %

**N = begin major heading

NP = begin major division without heading

**P = end major heading

**Q = " (begin quotation)

**R = begin minor heading

RT = begin minor division without heading

**S = ;

**T = end minor heading

*/ = begin Roman numeral

*, = end Roman numeral

*+0 = ° (degree)

**- = -- (dash)

**= = begin bold face

**$ = end bold face

**(= begin capitalization

**) = end capitalization

**. = abbreviation period

**.. = abbreviation period at end of sentence

With the exception of **Q **N **R (*(**(**/ and sometimes certain other marks as described under *Quotations* below, marks of punctuation are followed by space. **A is not followed by space when it is an internal apostrophe: thus can't is transliterated CAN**AT. Period is not followed by space when it is a decimal point. **D follows the letter over which the diaeresis or umlaut symbol appears; thus naïve is transliterated NAI**DVE. An ellipsis at the end of a sentence, indicated in the text by four periods, is indicated by **H followed by period without intervening space. (An ellipsis is counted as a single word in the word count). Where the text has only three suspension points at the end of a sentence, **H is used without following period.

A mark of punctuation following a passage in italics or bold-face is put after the close special type marker unless it clearly belongs to the special type. This gives the following transliterations:

box.	BOX*$.	*box*-	BOX*$-
box!	BOX**X*$	*box* !	BOX*$**X

box,	BOX,*$	*box,*	BOX*$,
box"	BOX**U*$	*box"*	BOX*$**U

Quotations. The Patent Office procedure of placing all punctuation marks after the close-quote symbol was followed. This leads to ambiguity, in that the distinction between an exclamation point or question mark inside a quotation and one outside it is lost. Thus the two sentences:

You said, "He's coming?"

You said, "He's coming"?

will both be transcribed *YOU SAID, **Q*HE**AS COMING**U**I This is one occasion where it would have been wiser to depart from the Patent Office code. An exception has been made when the punctuation mark belongs to a special type but the end-quote mark does not. Thus

You said, "He's *coming!*"

will be transcribed *YOU SAID, **Q*HE**AS *=COMING**X*$**U in contrast to the following:

You said, "He's *coming!*"	*YOU SAID, **Q*HE**AS *=COMING**U**X*$
You said, "He's *coming*!"	*YOU SAID, **Q*HE**AS *=COMING*$**U**X

The same applies to single quotes, indicated by **A as both open and close symbol. Some inconsistencies may have got into the text in this regard.

"Blocked" quotations, indicated by lefthand indentation, smaller type, or both, are treated as if enclosed within quotation marks. This includes also blocked quotations of verse, whose irregular line endings have not been marked.

Hyphenation. Unambiguous hyphens in the text have been preserved. Hyphens at the ends of lines of the original text are ambiguous, since they may be meant to indicate hyphenated words, in which case they are to

be preserved, or merely broken words, in which case they are to be ignored. The following practices were followed with line-end hyphens:

1. Those which clearly indicate broken words — *i.e.* which appear at a position which cannot be the dividing point between two parts of a compound word — have been ignored.

2. Those which occur at a position marking a possible division of a hyphenated compound word have been preserved if the compound occurs with an unambiguous hyphen elsewhere in the portion of text sampled.

3. In all other cases an arbitrary decision to preserve or omit the hyphen was made, based on the general practice of the text and the listings in Webster's *New International Dictionary, Third Edition.* All such arbitrary decisions are recorded in the detailed listings later in this manual.

Typographical Errors and Inconsistencies. No alterations have been made in the original text, even in the case of obvious typographical errors, misspellings, typographical inconsistencies, etc. All such errors that were observed have been recorded in the detailed listings. Users of the Corpus detecting errors not so recorded are urged to report them to the authors at the Department of Linguistics, Box E, Brown University, Providence R.I. 02912 so that future copies of the tape or this manual may be corrected.

4. THE TAGGED VERSION

In the tagged version of the Corpus (Form C), each individual word is furnished with a brief tag which assigns it to a specific word-class. There are 81 of these tags, which are listed on pages 23-25. They are of

six kinds:

(a) major form-classes ("parts of speech"): noun, common and proper; verb; adjective; adverb; in short, the open lexical classes;

(b) function words: determiners, prepositions, conjunctions, pronouns, etc.; the closed lexical and grammatical classes;

(c) certain important individual words: *not*, existential *there*, infinitival *to*, the forms of the verbs *do*, *be*, and *have*, whether auxiliaries or full verbs;

(d) punctuation marks of syntactic significance;

(e) inflectional morphemes, notably noun plural and possessive; verb past, present and past participle, and 3rd singular concord marker; comparative and superlative adjective and adverb suffixes. Thus when the following symbols appear elsewhere than in the first two letters of a tag, they normally have the following equivalences:

S = plural	D = past tense
$ = possessive	Z = 3rd singular verb
R = comparative	N = past participle
T = superlative	G = present participle or gerund
O = objective case of pronoun	

(f) two tags, FW and NC, are hyphenated to the regular tags to indicate that a word is a foreign word or a cited word, respectively. Thus the word *de* tagged FW-IN indicates that it is a foreign preposition. In a sentence such as "The word *if* has two letters," *if* would be tagged CS-NC.

Since the purpose of the tagged corpus is to facilitate automatic or semi-automatic syntactic analysis, the rationale of the tagging

system is basically syntactic, though some morphological distinctions with little or no syntactic significance have also been recognized. On the whole, the taxonomy is traditional and should be transparent to the grammarian, but in some areas distinctions have been made that may not be immediately obvious. They will be briefly explained here. The full rationale for the system is presented in *Automatic Grammatical Tagging of English*, by Barbara B. Greene and Gerald M. Rubin (Providence: Brown Univ., 1971.

The Noun Phrase. The model for this consists of a head preceded by a determiner sector and a modifier sector. The center of the determiner sector is the determiner itself, of which three basic kinds are recognized: articles, *a/an* and *the*, tagged AT; deictics, *this, that, another each*, tagged DT, with the plurals *these, those* tagged DTS, and the duals *either, neither*, which may also function as part of correlative conjunctions, tagged DTX; quantifiers *some, any*, not marked for number, tagged DTI. Preceding the determiner are the pre-quantifiers *all, half*, tagged ABN, and *both*, tagged ABX since it may also be part of a correlative conjunction. Between the determiner and the modifier position various elements may appear: a set of post-determiners, tagged AP, mostly but not all quantifiers such as *many, more, most, several, single*, with some particularizers such as *past, next, same, only*; cardinal and ordinal numerals, tagged CD and OD respectively; possessive nouns and pronouns, all having tags ending in $. The modifier sector may have positive, comparative, or superlative adjectives, tagged JJ, JJR, and JJT, and present and past participles, tagged VBG and VBN. Adjectives may be modified by qualifiers, such as *rather, very, too*, tagged QL, or followed by the post-qualifiers *enough, indeed*, tagged QLP. In general, adverbs

in *-ly* have not been tagged QL even when they serve a qualifying function; they are given the general adverb tag RB. There are, however, three exceptions — *awfully, fairly,* and *really.* Thus *a very hot day* or *a fairly hot day* is tagged AT QL JJ NN, but *an extremely hot day* is tagged AT RB JJ NN.

Certain adjectives which are semantically superlative and thus never compared are given the tag JJS; examples are *chief, head, main, prime, principal, single, top.*

A difficult taxonomic problem is posed by the fact that in English a large variety of words may appear as noun-modifiers between the determiner sector and the noun head. The tagging system makes provision for three kinds: adjectives, participles, and nominals. The problem lies in the fact that by compounding (open, hyphened, or closed), suffixation, or simple adjunction, English permits a bewildering variety of words, many of them nonce-constructions, to fill this position. Especially difficult are words or compounds with the suffixes *-ed* and *-ing* and/or the prefix *un-*. The following rules have been followed:

(1) hyphenated words which are legitimate noun phrases without the hyphen have been tagged NN (noun); thus *long-range, high-energy;*

(2) single words or compounds ending in *-ed* or *-ing* which are bona fide verbs when the ending is removed have been tagged VBN (past participle) or VBG (present participle) respectively; thus *untied, downgraded, outdistancing, double-crossing.* The exception is that when one of these words is modified by a qualifier, it is tagged JJ (adjective), as in *very tired, rather entertaining.*

(3) words normally nouns appearing in the immediate prenominal

position are treated as noun-adjuncts and tagged NN; thus *army officer, weather report*;

(4) all other words in the modifier position are tagged JJ. This means that the class of adjectives is the residual class, and is thus a large and anomalous one. Among the curiosities it includes are:

Words ending in -TYPE: SANDWICH-TYPE

Noun-Adj. combinations: FANCY-FREE, SCREW-LOOSE, SHOULDER-HIGH

Noun-pres. part. constructions: RUN-SCORING, SALES-BUILDING, LAW-ABIDING

Noun-past part. constructions: HOME-MADE, ROCK-STREWN (both of these also occur solid in the Corpus)

Noun-noun+-ED combinations: SHIRT-SLEEVED

Adj.-noun+-ED combinations: SHORT-SKIRTED, SLIM-WAISTED

Miscellaneous combinations: SHOW-OFFY, SIGNAL-TO-NOISE, SMASH-'EM-DOWN (modifying ADVENTURES), SNOB-CLANNISH, TOPSY-TURVY, TO-THE-DEATH, TONGUE-IN-CHEEK, TOO-SIMPLE-TO-BE-TRUE, UNIQUE-INGROWN-SCREWEDUP, ROUND-THE-CLOCK, DAY-AFTER-DAY

Both HIGH-POWER and HIGH-POWERED occur in the Corpus; the procedures outlined above tag the first of these NN and the second JJ.

In the head position, common nouns are tagged NN, with NNS used for the plural and NN$, NNS$ for possessive forms. Capitalized nouns are considered proper nouns and are tagged NP. Names and titles consisting of more than one word are tagged NP throughout, regardless of the word-class of the constituent items; thus *The Way of the World* is tagged NP NP NP NP. A block of NP tags without internal punctuation is thus normally a signal of a single name or title, which can be syntactically treated as a single word, usually a noun. Practice here has not been

entirely consistent; there may be instances of phrasal names in which non-capitalized constituents are given their regular tags instead of NP. These will be corrected as they are discovered.

The Verbal Phrase. Verbs in the base form, regardless of syntactic function, are tagged VB. The inflected forms of normal verbs are marked with the suffix tags Z (3rd. singular), D (past tense), N (past participle), and G (present participle/gerund). Modal auxiliaries, regardless of tense, are all tagged MD. The verbs *be, have,* and *do,* whether serving as auxiliaries or as full verbs, have the special tags BE, HV, and DO, with inflectional variants (exceptions: *doing* and *done* are tagged VBG and VBN). This permits the ready identification and analysis of verbal phrases. The archaic forms *art* and *hast* are tagged as base forms, while *hath* and other forms in *-th* are tagged as 3rd singular (HVZ, VBZ). Contracted forms of auxiliaries have their regular tags joined to the subject tag by +; thus *I'm, you've,* and *he'll* are tagged PPSS+BEM, PPSS+HV, and PPS+MD respectively. Contracted negatives have the tag for *not*, *, immediately following the verb tag; thus *can't* is tagged MD*. Condensed forms in dialogue, such as *gonna*, are tagged for their morphemic constituency, VBG+TO.

Pronouns. Personal pronouns have tags beginning with PP, followed by one or more letters indicating case, concord, and sometimes number. All subject forms which concord with the base form of the verb in the present are tagged PPSS, regardless of person and number. Those that concord with the *-s* form of the present are tagged PPS. Forms in object function, whether or not morphologically marked, are tagged PPO. First possessives (e.g. *my, our*) are tagged PP$; second (nominal) possessives

are tagged PP$$ (*mine, our*). Reflexive/intensive pronouns are tagged PPL if singular and PPLS if plural, with no distinction for case. Interrogatives and relatives begin with WP; subject forms are tagged WPS and object forms WPO.

Indefinite pronouns — compounds of *any-, every-, no-,* and *some-* — are tagged PN, or PN$ if they have the possessive suffix *-'s*.

So-called demonstrative pronouns — *this, that,* etc. — are treated as free-standing determiners and tagged accordingly: DT, DTI, DTS.

Adverbials. The general tag for adverbs is RB, with RBR and RBT for inflectional comparatives and superlatives. The case of qualifiers like *very, fairly* is discussed above under the Noun Phrase. Certain adverbs, mostly temporal or locative, which often function as nominals have been denominated "nominal adverbs" and tagged RN; thus *here, then, indoors.* Conversely, locative and temporal nouns which often function adverbially have been denominated "adverbial nouns" and tagged NR; thus *home, east, Tuesday.*

In the case of phrasal or two-part verbs, the attempt was at first made to distinguish between adverbs (*hold out your hand*) and particles (*hold out for more money*). It was found, however, that this necessitated a large number of arbitrary decisions, which might confuse or mislead those using the tagged Corpus. It was decided instead to consider this a syntactic and semantic rather than a taxonomic problem, and to give the "portmanteau" tag RP (for "adverb or particle") to the ten words *about, across, down, in, off, on, out, over, through,* and *up,* except when they are functioning as prepositions, when they receive the normal preposition tag IN.

Connectives. Coordinating conjunctions (*and, or,* etc.) are tagged CC and subordinators (*since, because, if*) CS. Prepositions are tagged IN. The word *to* is tagged TO when used as the infinitive marker.

Miscellaneous Items. The existential subject *there* is tagged EX, and thus distinguished from the homonymous adverb. Exclamations of various sorts, which have no syntactic function, are tagged UH; they occur mostly in the dialogue of the fictional samples. The word *not* is tagged *, which is joined to the verb tag in the case of contracted forms. The use of the foreign word tag FW and the metalinguistic citation tag NC has been explained above.

***

The tagging of the Corpus has been a long and arduous process, extending over several years and involving quite a few different people. Although elaborate proofreading and checking procedures have been used, it is inevitable that errors and inconsistencies remain. Users of the tagged Corpus detecting such errors and inconsistencies are urged to let us know so that they may be corrected. Lists of corrigenda will be sent from time to time to all holders of the tagged Corpus. Please address corrections and suggestions to the authors at Text Research, 196 Bowen Street, Providence RI 02906.

Capitalized Words, Titles, and Proper Nouns

The conventions of (non-sentence-initial) capitalization in English are complex and to a considerable degree variable, unlike most other aspects of the writing system. This has presented a problem in tagging, which has been disposed of, if not settled, by arbitrary rules. These have been made as objective as possible, but there still remain cases where judgment is necessary and hence inconsistency is possible. The aim has been to identify capitalized uses as much as possible, but also to associate capitalized words with their lower-case alternatives. This has been done by means of the tags NP (with its inflected variants NP$, NPS, and NPS$) and TL. The former of these is a primary tag and the latter a tag hyphenated to other primary tags. The following procedures have been observed.

1. Sentence-initial capitals have been reduced to lower case for all words except those identified as bearing capitals when not initial.

2. Random capitalization occurring in quotations from older sources or illiterate writers has been preserved, but without recognition by tag. The following sentence from Sample G38, a quotation from Defoe, illustrates eighteenth century practice:

> If they Could Draw that young Gentleman into Their Measures They would show themselves quickly, for they are not asham'd to Say They want Onely a head to Make a beginning.

The capitalized words in passages of this sort have been given their normal tags, but the spellings are preserved as variants in the lemmatized list.

3. The same practice has been followed with German nouns, commonly capitalized. They receive the tag FW-NN.

4. Words identifiable as "proper nouns," that is, nouns used to designate particular individuals, types, or groups, without further semantic content, are tagged NP (NPS for plurals; NP$ and NPS$ for singular and plural possessives). Examples:

John F. Kennedy's daughter Caroline
NP NP NP$ NN NP

5. Words occurring as constituents of titles, e.g. of books, plays, corporations, government agencies, etc., are given their normal tag with the addition of the hyphenated tag -TL. In most cases these words are capitalized, except for function-words such as prepositions, conjunctions, and sometimes pronouns. Some examples:

the United States of America

VBN-TL NNS-TL IN-TL NP-TL

Gulliver's Travels

NP$-TL NNS-TL

the Protestant Episcopal Church

JJ-TL JJ-TL NN-TL

It is to be noted that in some languages--French, for example--words in titles are often not capitalized:

Recherches sur l'identité des forces chimiques

FW-NNS-TL FW-IN-TL FW-AT+NN-TL FW-IN+AT-TL FW-NNS-TL FW-JJ-TL

et électriques

FW-CC-TL FW-JJ-TL

6. A problem is presented by names of persons and places which are homographs (and often cognates) of common words. Somewhat arbitrarily, the following procedures have been followed:

a. Names of persons have all been tagged NP, regardless of their etymological status:

Gen. Thomas Power

NN-TL NP NP

cp. Georgia Power Company

NP-TL NN-TL NN-TL

b. Geographical terms and other descriptive words forming parts of place-names are given their basic tags followed by -TL

the Allegheny Mountains

AT NP-TL NNS-TL

the Great Smoky Mountains

AT JJ-TL JJ-TL NNS-TL

c. Proper names forming parts of place names are tagged NP-TL

See "Allegheny Mountains" above

d. The titles Mr., Mrs., and Miss have been tagged NP.

e. Other titles of persons which have distribution as common nouns, adjectives, etc., have been given their regular tags plus -TL:

Mayor William B. Hatfield

NN-TL NP NP NP

Secretary General Dag Hammarskjold

NN-TL JJ-TL NP NP

g. Foreign and/or exotic titles not appearing as common nouns, etc., in the Corpus have been tagged NP:

Signora Ferraro

NP NP

LIST OF TAGS

Tag	Description	Examples
.	sentence closer	. ; ? !
(	left paren	
)	right paren	
*	*not, n't*	
--	dash	
,	comma	
:	colon	
ABL	pre-qualifier	*quite, rather*
ABN	pre-quantifier	*half, all*
ABX	pre-quantifier	*both*
AP	post-determiner	*many, several, next*
AT	article	*a, the, no*
BE	*be*	
BED	*were*	
BEDZ	*was*	
BEG	*being*	
BEM	*am*	
BEN	*been*	
BER	*are, art*	
BEZ	*is*	
CC	coordinating conjunction	*and, or*
CD	cardinal numeral	*one, two, 2, etc.*
CS	subordinating conjunction	*if, although*
DO	*do*	
DOD	*did*	
DOZ	*does*	
DT	singular determiner	*this, that*
DTI	singular or plural determiner/quantifier	*some, any*
DTS	plural determiner	*these, those*
DTX	determiner/double conjunction	*either*
EX	existentil *there*	
FW	foreign word (hyphenated before regular tag)	
HV	*have*	

Tag	Description	Examples
HVD	*had* (past tense)	
HVG	*having*	
HVN	*had* (past participle)	
IN	preposition	
JJ	adjective	
JJR	comparative adjective	
JJS	semantically superlative adjective	*chief, top*
JJT	morphologically superlative adjective	*biggest*
MD	modal auxiliary	*can, should, will*
NC	cited word (hyphenated after regular tag)	
NN	singular or mass noun	
NN$	possessive singular noun	
NNS	plural noun	
NNS$	possessive plural noun	
NP	proper noun or part of name phrase	
NP$	possessive proper noun	
NPS$	possessive plural proper noun	
NR	adverbial noun	*home, today, west*
OD	ordinal numeral	*first, 2nd*
PN	nominal pronoun	*everybody, nothing*
PN$	possessive nominal pronoun	
PP$	possessive personal pronoun	*my, our*
PP$$	second (nominal) possessive pronoun	*mine, ours*
PPL	singular reflexive/intensive personal pronoun	*myself*
PPLS	plural reflexive/intensive personal pronoun	*ourselves*
PPO	objective personal pronoun	*me, him, it, them*
PPS	3rd. singular nominative pronoun	*he, she, it, one*
PPSS	other nominative personal pronoun	*I, we, they, you*
QL	qualifier	*very, fairly*
QLP	post-qualifier	*enough, indeed*
RB	adverb	
RBR	comparative adverb	
RBT	superlative adverb	

Tag	Description	Examples
RN	nominal adverb	*here, then, indoors*
RP	adverb/particle	*about, off, up*
TO	infinitive marker *to*	
UH	interjection, exclamation	
VB	verb, base form	
VBD	verb, past tense	
VBG	verb, present participle/gerund	
VBN	verb, past participle	
VBZ	verb, 3rd. singular present	
WDT	*wh-* determiner	*what, which*
WP$	possessive *wh-* pronoun	*whose*
WPO	objective *wh-* pronoun	*whom, which, that*
WPS	nominative *wh-* pronoun	*who, which, that*
WQL	*wh-* qualifier	*how*
WRB	*wh-* adverb	*how, where, when*

NOTES

(1) Merged constructions are marked by the appropriate tags joined by +, except for *, which is affixed directly. For example:

isn't BEZ* *he'd* PPS+HVD or PPS+MD

there's EX+BEZ or EX+HVZ or RN+BEZ

(2) The tags FW, NC, NP are given to single words, phrases, or sentences contained respectively in foreign phrases, cited passages, and compound or complex titles or names. The first two are hyphenated to the regular tag. For example:

mens	FW-NN	*in*	IN
sana	FW-JJ	*drug*	NN-NC
in	FW-IN	*store*	NN-NC
corpore	FW-NN	*the*	AT
sano	FW-JJ	*primary*	JJ
		stress	NN
		is	BEZ
		on	IN
		drug	NN-NC

5. COPYRIGHT RESTRICTIONS

Forms A and B, Bergen Types I and II, Stanford Brown MARC Form.

Most of the selections from which the samples were chosen are under copyright. Copyright holders have generously permitted their use without payment of fee, with the understanding that the Corpus is to be used primarily for scholarly research in linguistics, stylistics, and other relevant disciplines. The following restrictions apply to all copies of the Corpus except Form C, which is discussed below; persons or institutions requesting copies of the tapes will be asked to subscribe to them before copies will be issued.

1. No copies of the tapes are to be made for any use except within the institution holding the tapes without the written permission of the Department of Linguistics at Brown University.

2. Print-outs of the Corpus or parts thereof are to be used only for bona fide research of a non-profit nature. Holders of copies of the Corpus tapes may not reproduce any samples or parts of samples other than short extracts considered to come under "fair use" provisions without getting written permission of the individual copyright holders, as listed in the detailed description of the individual samples later in this manual.

3. Commercial publishers and other non-academic organizations wishing to make public use of part or all of the Corpus must obtain permission from the Department of Linguistics, Brown University. They may be asked to get written permission from individual copyright holders.

Form C.

The tagged Corpus is copyrighted in its entirety by W. N. Francis and Henry Kučera. Permission of one of the copyright holders must be

obtained before any part of the tagged Corpus is reproduced in any form.

6. BASIC TECHNICAL INFORMATION

As described in Section 2 above, the Corpus is available in several formats. Forms A, B, and C are available from the authors.

Forms A and B contain only the text. Form C contains both the text and the grammatical annotation, as described in Section 4 above.

The only difference between Forms A and B is that Form A includes all of the graphic coding described in Section 3 above, while all this graphic coding except hyphens, apostrophes, and symbols for formulas and ellipses has been removed from Form B. In all other respects the formating of Forms A and B is identical. The format of the tagged Form C, on the other hand, is quite different. It is described in detail later in this Section.

Forms A and B: The Corpus Text.

The organization of the text data on tape corresponds to the punched format described above. The data is recorded in card-image form, i.e. all cards (including correction cards and other incompletely filled cards) are reproduced in their entirety. Each logical record on the tape thus consists of 80 characters, of which the first 70 represent the text proper and the last 10 are reserved for the location marker. The structure of the location marker in Forms A and B is that given in Section 3 above. (The user should be cautioned that the structure of the location marker is different in the tagged Form C, described below, where the location marker contains more information).

In the standard versions of the Corpus, the 80-character records are blocked by a factor of 40, so that the blocksize is 3200 characters. Users desiring copies of the tape in Form A or B with a different blocksize should make special arrangements with the authors.

Forms A and B are available in upper-case characters only, in either ASCII or EBCDIC coding, on 9-channel tapes in the following densities: 800, 1600, and 6250 BPI. (7-channel tapes are available by special arrangement only). The tapes have an IBM standard label. Users who do not wish the label included should make a specific request to this effect.

Since the Corpus contains 500 separate samples, it was considered desirable to indicate the end of each sample by easily discernible means. This is accomplished by signaling the end of each sample by the insertion of 69 or more blanks. It should be emphasized (as already specified above) that less than 69 blanks do not indicate the end of a sample but are equivalent to a single blank in the coding system (see p. 9 above). Since sample size and tape record length are independent of each other, the end of a sample does not necessarily coincide with the end of a tape record.

Form C, the Tagged Version.

The tagged version of the Corpus consists of 1,136,857 fixed-length records,[7] with each record 52 characters long. In the normal format of Form C, these records are blocked by a factor of 100, so that each block consists of 5200 characters. In this format, two 2400-foot tapes

[7]This number is considerably larger than the total number of words given on p. 1 above because certain external punctuation marks (see below) are treated as separate "words."

are needed to accommodate the entire tagged Corpus. Users desiring copies of the tapes with a different blocksize should make special arrangements with the authors.

Form C is available in either ASCII or EBCDIC coding, on 9-channel tapes in the following densities: 800, 1600, and 6250 BPI (7-channel tapes are available by special arrangement only). The tapes have an IBM standard label. Users who do not wish the label included should make a special request to that effect.

Each of the 1,136,857 records contains three items of information: the word or external punctuation symbol; the grammatical tag; and the location marker which specifies the position of the word in the Corpus. These three items of information are given in fixed-length fields within each record, and the information is left-justified within each field. The layout of the fields comprising a record is as follows:

Columns 1-30 the word or external punctuation symbol

Columns 31-41 the grammatical tag

Columns 42-52 an eleven-character location marker.

Lexical Items (Columns 1-30) consist of graphic words or numerals, and are coded in upper-case ASCII or EBCDIC character. Capitalization is indicated only for proper names; the symbol for capitalization is an asterisk [*] immediately preceding the capitalized letter. No capitalization is given for sentence-initial letters. So, for example, the opening phrase of the Corpus, *The Fulton County Grand Jury said Friday* consists of seven records, whose lexical items are coded as follows:

THE

*FULTON

*COUNTY

*GRAND

*JURY

SAID

*FRIDAY

Aside from the symbol for capitalization, three symbols of internal-record punctuation are utilized: a hyphen, an apostrophe, and an abbreviation period. The abbreviation period, which (as explained below) is distinct from the sentential period, follows directly the last character of the abbreviated word. For example, the date *Aug. 31* consists of two records, and is coded as follows:

*AUG.

31

No lexical item given in columns 1-30 contains any internal blanks.

External Punctuation. There are nine symbols and six tags used for external punctuation. The symbols are period [.], exclamation mark [!], question mark [?], semicolon [;], comma [,], colon [:], dash [coded as two hyphens, --], left parenthesis [(], and right parenthesis [)]. Each occurrence of such external punctuation constitutes a separate record: the punctuation symbol is in column 1 (or columns 1 and 2 in the case of a dash). The tags for punctuation symbols are identical to the punctuation symbols themselves, except that period, exclamation mark, question mark, and semicolon, which are considered sentence delimiters, all have the same tag, [.]. There are thus six different tags for the nine punctuation symbols (see p. 23). The tag is in column 31 (columns 31 and 32 for dash), and the location marker, of the same structure as that for other lexical items (cf. below), is in columns 42-52.

Grammatical Tags (columns 31-41) begin in column 31 and consist of continuous strings of upper-case characters and certain special symbols, as listed on pages 23-25 above.

Location Markers (columns 42-52). Each location marker begins in column 42 and is eleven characters long. The first three positions identify the genre and sample number in which the word occurs. Genres, identified in the first position of the location marker by a single letter of the alphabet (A through R, omitting I, O, and Q; see pages 3-5 above for the full list) are the fifteen major categories of samples. Positions 2 and 3 of the location marker give the sample number within the genre; these are individually listed on pages 33-176 below. Positions 4-7 give the line number of the word's occurrence, according to the original numbering of Forms A and B, and positions 8 and 9 specify the sequential location of the word in that line. Thus, in contrast to Forms A and B, where the location extends only to the line of text in which a word occurs, in Form C each word has a unique location marker. So, for example, the first nine positions of the location marker A02103014 identify the word as occurring in genre A, sample 2, line 1030, and as the 14th word in that line.

The last two positions of the location marker are invariable and contain the symbols E1 (for English Corpus One); they are included simply to differentiate this data set from other grammatically analyzed corpora which may become available.

The following is an example of an actual record from Form C:

SAID VBD A01001006E1

This record identifies the word *said* as the past tense form of a verb and locates it in genre A, sample 1, line 1, as the sixth word in that line.

7. THE INDIVIDUAL SAMPLES

The following pages list the 500 samples of the Corpus individually. Each entry contains the following information:

Sample Number. Author (if known) and title.

Publication Information.

Copyright statement.

Number of Lines

Special Information: typographical errors and inconsistencies, arbitrary hyphens, and other relevant information about the text.

Word Count Number and percentage of words in quotes. Number of symbols and formulas.

In the case of samples drawing from more than one source (principally in the Press sections A-C) the lines occupied by each subsample are indicated.

A. PRESS: REPORTAGE

A01. *The Atlanta Constitution*

Used by permission of *The Atlanta Constitution*, Constitution State News Service (H), and Reg Murphy (E).

A.	November 4, 1961, p. 1	"Atlanta Primary..."	0010-0670
B.		"Hartsfield Files"	0680-0850
C.	August 17, 1961, p. 6	"Urged Strongly..."	0860-1060
D.		"Sam Caldwell Joins"	1070-1130
E.	March 6, 1961, p. 1	"Legislators Are Moving" by Reg Murphy	1140-1390
F.		"Legislator to Fight" by Richard Ashworth	1400-1530
G.		"House Due Bid..."	1530-1670
H.	p. 18	"Harry Miller Wins..."	1670-1920

Arbitrary Hyphen: multi-million [0520]

1,988 words 431 (21.7%) quotes 2 symbols

A02. *The Dallas Morning News*, February 17, 1961, section 1

Used by permission of *The Dallas Morning News*.

A.	p. 5	"Committee OK..." by Jimmy Banks	0010-0360
B.		"Austin Wire..." by Dawson Duncan	0370-0990
C.		"Legislators Reject..."	1000-1140
D.		"Water Development" by Richard M. Morehead	1150-1360
E.		"Cut Proposed..."	1370-1520
F.	p. 12	"Paris College..."	1530-1700
G.		"Principals..."	1710-1740

R05, continued.

Used by permission of Evan Esar. 0010-1850

Typographical Error: indefinity [1400]
Arbitrary Hyphen: girl-friend [0230] No Hyphen: standby [0500]

2,024 words 839 (41.5%) quotes 1 formula

R06. James Thurber, "The Future, If Any, of Comedy," *Harper's Magazine* 223: 1339 (December, 1961), 40 - 43.

Used by permission of Mrs. James Thurber. 0010-1780

2,015 words 1,646 (81.7%) quotes

R07. John Hazard Wildman, "Take It Off," *The Arizona Quarterly*, 17: 3, (Autumn, 1961), 246 - 252.

Used by permission of *The Arizona Quarterly*. 0010-1770

2,028 words 356 (17.6%) quotes

R08. Leo Lemon, "Catch Up With" and "Something to Talk About," *Mademoiselle*, July, 1961, 8, 15, 17, 47 - 49.

Used by permission of *Mademoiselle*. 0010-1910

Arbitrary Hyphens: tongue-twister [0380] biri-pitknoumen [0390]

Note: F-major [0860] N minor [0870]

Typographical Error: unwaivering [1810]

2,027 words 118 (5.8%) quotes 5 symbols

R09. S. J. Perelman, *The Rising Gorge*. New York: Simon and Schuster, 1961. Used by permission of S. J. Perelman. Pp. 201 - 207.

2,004 words 47 (2.3%) quotes 1 symbol 0010-1740

LIST OF SAMPLES

LIST OF SAMPLES

LIST OF SAMPLES

LIST OF SAMPLES