DICTIONARY OF ENGLISH ETYMOLOGY FOR NLP AND RELATED TAGGING UTILITY PROGRAMS

SARAKI Masashi Saraki R&D
OSADA Tetsuo Waseda University
WATANABE Yuhki Toyohashi University of Technology
SHIRAI Satoshi NTT CS Laboratories

1 INTRODUCTION

Etymology is the study of the origin of words. Etymology is a remarkable attribute of words, indicates as it does from where words have come from or have been borrowed. In this sense, the etymology never changes with the times and, in fact, has never changed throughout the history of languages. The signification of individual words, however, has changed and will be able to continue doing so with the times. Setting up etymology as one attribute of any word, we can create the potential for another tool for syntactical and rhetorical analysis.

The historical process of the development of English is reflected in its contemporary vocabulary and expressions. English vocabulary and expressions consist of three layers, just as in geologic stratification: a primal layer, an intermediate layer, and a modern layer, referred to as Saxon, French, and Latin respectively. The primal, Saxon layer is psycholinguistically the deepest in the mind of the native speaker and is tied to images of the soul, the French includes concepts of social consciousness and the Latin involves logos, or reason.

2 DICTIONARY OF ENGLISH ETYMOLOGY FOR NLP

The Dictionary of English Etymology for Natural Language Processing (hereafter simply referred to as ***DEE***) includes the origin and borrowing process of English words, but does not record semantic history, that is, the original meaning and the changes in significance the words have undergone. The historical approach enables the user of ***DEE*** to view the historical process of individual words. ***DEE*** has now approximately 20,000 words and will be further increased to 25,000.

2.1 Structure and Content of Each Entry

Referring to **Figure 1**, an entry consists of items which appear in the following order, although except serial number and head word, not all of these will necessarily be found in any particular entries:

- Serial number
- Head word (A)
- Word class (B)
- Chronology (C, D)
- Etymology (F, G, ...)

Figure 1 ***DEE Database***

Microsoft Excel - ?
ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

	A	B	C	D	E	F	G
4499	compensate	v.	1656	ModE		L.	compensare
4500	compensation	n.	1387	ME		L.	compensationem
4501	compensatory	adj.	1601-2	ModE			
4502	compere	n.	1738	ModE		F	compere
4503	compere	v.	1933	PE			
4504	compete	v.	1541	ModE		late L.	competere
4505	competence	n.	1594	ModE		F	competence
4506	competent	adj.	1400	ME		ME	
4507	competition	n.	1608	ModE		L	competitionem
4508	competitive	adj.	1829	PE		L	competere
4509	competitor	n.	1534	ModE		F	competiteur
4510	compilation	n.	1430	ME		F	compilation
4511	compile	v.	1425	ME		ME	
4512	compiler	n.	1330	ME		OF	compileor
4513	complacence	n.	1430	ME		med L	complacentia

2.2 Chronology and Etymology

Chronology indicates the year of first usage and time divisions of English. The marking for Etymology indicates the particular foreign language in which a word originated or from which it was borrowed. The etymological item mentions the origin and word formation of the head word, and if a loan word, further tracing its historical process of the word back to its true origin.

2.3 Structure of the DEE database

The ***DEE*** database includes a main database as shown in Figure 1 and supplementary databases of irregular verbs, postposed adjectives and affixes. Additional databases of other word classes will be incorporated.

3 TAGGING UTILITY PROGRAMS

When used in conjunction with the ***DEE*** database, the tagging utility will allow users to make the maximum use of data. The utility allows users to tag subsequent etymological labels to each word of the user's text and thus to view etymological word arrangement.

3.1 A Set of Etymological Labels

Some examples are shown hereinafter:

OE ME, ModE , PE;[1]
F, OF, AF, NF[2];
L, late L, med L, mod L, VL;[3] Gk[4]
Ir, Ital., Sp, Ar[5];

3-2 Utility Programs

Utility includes a tagging program, a user interface, and database files of the etymology. The tagging program is implemented in using PERL and the user interface is available in CGI, as shown in **Figure 2**. The CGI interface allows the user to select a desired text and run the tagging program. The user can select optional labels by marking check boxes. Thus, the user can choose word classes such as Nouns, Verbs, Adjectives, Adverbs, Pronouns, Prepositions, Articles, Auxiliaries, Copula, Coordinates, Subordinates, and Interrogates to be tagged with the etymological labels.

3.3 The Result of The Tagging

The following original text is a excerpt of an English novelist's work[6] with the text tagged according selected labels, for example, nouns, verbs ,adjectives, adverbs, and subordinates.

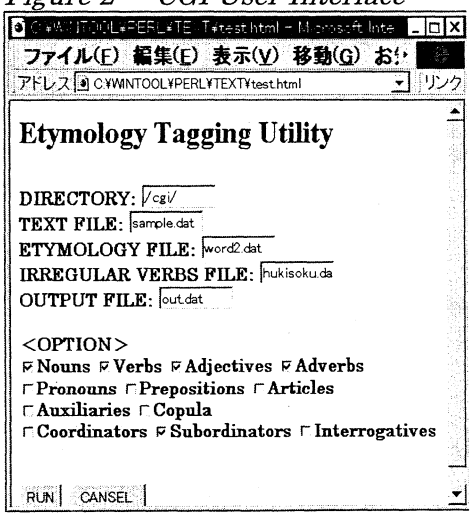
Original Text

"Most people who bother with the matter at all would admit that the English language is in a bad way, but it is generally assumed that we cannot by conscious action do anything about it. Our civilization is decadent and our language so the argument runs must inevitably share in the general collapse."

Tagged Text

"Most(OE) people(AF) who bother(ON) with the matter(OF) at all would admit(L) that the English(OE) language(OF) is in a bad(OE) way(OE), but it is generally(L-OE) assumed(L) that we cannot by conscious(L) action(F) do anything about it." Our civilization(F) is decadent(L) and our language(OF) so(OE) the argument(F) runs(OE) must inevitably(L) share(OE) in the general(L) collapse(L).

Figure 2 CGI User Interface



[1] Old English, -1100), Middle English, 1100-1500), Modern English, 1501-1800), Present English, 1801-)

[2] French, Old French, Anglo French, Norman French

[3] Latin, late Latin, medieval Latin, modern Latin(Neo Latin), Vulgar Latin

[4] Classical Greek

[5] Irish, Italian, Spanish, Aramaic

[6] *Politics and the English Language, George Orwell, 1946*

4 THE RESULTS OF EMPIRICAL EXPERIMENTATION

4.1 Roget Thesaurus

Peter Mark Roget proposed a methodology for compiling his thesaurus in the introduction to the original edition in 1852[6]. The principle for Roget's classification is the same as that which is employed in the various branches of Natural History, and thus the sectional divisions Roget formed, corresponding to natural families in botany and zoology, and the filiation of words presents a network analogous to the natural filiation of plants or animals. Thus, Roget established "tabular synopsis of categories" and accordingly classified English vocabulary into six primary classes with further subdivisions. Words are arranged under several topics or head of signification. A portion of Roget's thesaurus, which has been revised, is cited below (the words have been tagged with the etymological labels):

Existence.(OF)

N. ***existence***(OF) being(OE) entity(medL); absolute(F) being(OE), the absolute(F) 965 ***diviness***(L); aseity(medL), self(OE)-exisatence(OF); monad(lateL), a being(OE), an entity(medL), ens(OF) essence(L), quiddity(medL); Platonic(L) idea(Gk), universal(OF); subsistence(lateL) 360 ***life***(OE); survival(AF), eternity(OF), 115 ***perpetuity***(OF); preexistence(OF) 119 ***priority***(OF) ; this life(OE) 121 ***present***(OF) ***time***(OE); existence(OF) in space(OF), prevalence(F) 189 ***presence***(OF); entelechy(lateL), realization(F), becoming(OE), evolution(L) 147 ***conversion***(OF); creation(OF) 164 ***production***(OF); potentiality(medL) 469 ***possibility***(MF); ontology(modL), metaphysics(Gk); realism(F), materialism(modL), idealism(F), existentialism(D) 449 ***philosophy***(L)

reality(OF), realness(OF-OE[7]), actuality,(medL) entelechy(lateL), Dasein(D); actual(OF) existence(OF), material(lateL) e. 319 ***materiality***l(OF-OF); thatness 80 ***speciality***(OF); positiveness(OF-OE); historicity(L-OF), factuality(L-OF), factualness(L-OF-OE) 494 truth(OE); fact(L), fact(L) of life(OE), undeniable(OE-ME-OF) f., positive(OF) f., stubborn(OE) f., matter(OF) of f., fait accompli (F)154 event(L); real(AF) thing(OE), not a dream(OE), no joke(L); realities(OF), nitty-gritty(America), basics(PE), fundamentals(modL), bedrock, nuts(OE) and(OE) bolts,(OE) brass(OE) tacks(AF) 638 important(medL) matter(OF)

essence(L), nature,(natura<nasci) very(OF) n., essential(medL) n., quiddity(medL), hypostasis(Gk) 3 substance(L); constitutive(L-OF) principle(OF), inner(OE) being(OE), sum(L) and substance(L) 5 essential(medL) part(L); prime(OF) constituent(L), soul(OE), heart(OE), core(OF), centre(L) 224 interiority(L-OF).

Further, by tagging the whole of Roget's thesaurus, we will be able to find the etymological characteristics of English words expressing general ideas

Secondly, Roget mentioned the notion of **correlative words**: "For the purpose of exhibiting with opposite and correlative ideas, I have, whenever the subject admitted of such an arrangement, placed them in two parallel columns in the same page, so that each group of expressions may be readily contrasted with those which occupy the adjacent column, and constitute their anthithesis."[6] Roget also suggested a method of arranging correlative terms in the form of a triad as follows:

α Two ideas which are completely opposed to each other, admitting of an intermediate or neutral ideas. In the following examples, the words in the first and third columns express opposite ideas, and the terms in the middle column have a neutral sense with reference to the former.

Identity(L)	Difference(OF)	Contrariety(OF)
Beginning(OE)	Middle(OE)	End(OE)

β The intermediate word is simply the negative to each of two opposite positions:

Convexity(L)	Flatness(ON[8])	Concavity(L)
Desire(OF)	Indifference(L)	Aversion(L)

γ The intermediate word is properly the standard with which each of the extremes is compared:

Insufficiency(lateL)	Sufficiency(L)	Redundance(L)

The same word has several correlative words according to the different relations:

Giving(ON)	Receiving(ONF)/ Taking(OE)
Old(OE)	New(OE)/ Young(OE)
Attack(F)	Defence(OF)/ Resistance(OF)
Resistance(OF)	Attack(F)/ Submission(OF)
Truth(OE)	Error(OF)/ Falsehood(OE)

[7]In cases in which the word formation itself has no origin, the label immediately before the first hyphen indicates the origin of a prefix, the label immediately following the last hyphen indicates that of a suffix, and an intermediate label between the labels indicates the origin of a root.

[8] from Old Norse flatr; akin to Old High German flaz flat, and probably Greek platys broad

Use(**OF**)	Disuse(**OF**)/ Misuse(**OF**)
Teaching(**OE**)	Misteaching(OE?)/ Learning(**OE**)

The origins in each triad are represented with a set of etymological labels. According to the tagged triads, we can work on the presumption that correlative ideas are expressed with corresponding correlative words having the similar origin.

4.2 Postmodification By Adjectives

Postposed adjectives have a prepositional phrase or nonfinite verb phrase as complement. English syntax allows adjectives both to be prepositive and postpositive in a noun phrase and further to be grouped immediately before a noun. This unique syntactical characteristic is an intermediate step between Roman and Germanic aspects. By tagging the etymological labels to postposed adjectives, we can show the data in tabular fashion, as shown in **Table 1** and a concordance in **Table 2**. We can presume that postmodification by adjectives was devised according to Latin grammar and derive the hypothesis that "If S+P+N1(French/Latin) +Adj.(French/Latin)+prep., then "Adj.+prep. is predicate of N1."

5 APPLICATION PERSPECTIVE

When ***DEE*** and its utility can work in conjunction with text processing, for example ***WRAPL***[7][8] which can convert complex English sentences to simple sentences, it will help widen access to English texts throughout the World Wide Web for people to whom English is a foreign language and assist people who have language disabilities to read with greater ease. ***DEE*** and ***WRAPL*** will be simplify English complex sentences into simple sentences and replace words that are abstruse or abstract words(loan words) with ones that are plain or concrete(native words). Thus, we will intend to contribute PSET(Practical Simplification of English Texts)[9] .

REFERRENCES

[1]*The Oxford English Dictionary, second edition*, 1989, Oxford University Press

[2]Onions, C.T, *Oxford Dictionary of English Etymology*, Oxford University Press

[3]*The Webster's Third International new Dictionary*, 1986, Merrian Webster.

[4]*The American Heritage Dictionary of the English language*, 1992, Houghton Mifflin

[5]*The Kenkyuusha Dictionary of English Etymology*, 1997, Kenkyuusya

[6]Roget, P.M., *Introduction to the original edition, 1852: Rogeh' Original Thesaurus of English words and phrases*,1995, pp.xviii -xxix, Viking

[7] Saraki, Kato, Ogawa, "Text Processing for Machine Translation by using WRAPL", Proceedings of The Third Annual Meeting of The Assocation for NLP, 1997, pp. 561-564

[8]Kato, Ogawa, Saraki "Pattern Analyzing of English Complex Sentences and Reducing Hypotaxis to Parataxis", Technical Report of IEICE, NLC971-17, 1997-07, pp. 65-70

[9] Jhon Tait et al., *"PSET: Description of Proposed Research"* http://osiris.sunderland.ac.uk/~pset/pset-prop-shortened.html http://www.cogs.susx.ac.uk/lab/nlp/pset/pset.html

Table 1

able	ME < OF < L
acceptable	ME <OF
accessible	1610 < OF < L
adjacent	ME < L
analogous	L < Gk.
apart	ME < OF < L
apparent	ME < OF
desirable	ME < OF < L
different	ME < OF < L
easy	ME < OF < L

enough	ME < OE
equal	L
equivalent	ME < LL
essential	ME < late L
Inferior	ME < L
inherent	1578< L
liable	AN < OF < L
movable	ME < OF
necessary	ME < L
opposite	ME < OF < L

payable (pay)	ME < OF < late L< L
perpendicular	ME < OF < L
placeable	ME < OF < L
reluctant	1667 < L
responsible	1599 < F < L
separate	ME < L
similar	1626 < F < L
superior	ME < OF < L
vulnerable	1609 < late L <
worthy	ME < OE

Table 2

useful(OF-OE), the contradiction(L) inherent(L) in the principle(OF) is irremediable(L-AF-L).
that strange(OF) contradiction(L) inherent(L) in human(OF) nature(OF), the Jekyll and Hyde elements(L) which,
arguments(OF) as to the dangers(OF) inherent(L) in the power(AF) of the State(L) have made(OE) them dissatisfied
end, under the disadvantages(OF) inherent(L) in this style(OF) of writing(OE).
by fomenting(lateL) the enmities(OF) inherent(L) in the nationalist(F) idea(lateL), and everywhere, after a very
horned(OE) monster(L) or to evil(OE) inherent(L) in the human(OF) heart(OE), Iwill not assume(L) to say(OE)
anthropological(Gk) factors(OF), inherent(L) in the individual(medL) criminal(F), are the first(OE) condition
this as with other frailties(OF) inherent(L) in our nature(OF); the desire(F) of deferring(L) to another
be the religious(L) instinct(L) inherent(L) in man(OE)--that perception(OF) so(OE) fine(OF), so subtle(OF)