The Dictionary of English Etymology for Analyzing Expressions

SARAKI Masashi OSADA Tetsuo NITTA Yoshihiko
saraki@st.rim.or.jp strawberry@pop06.odn.ne.jp nitta@eco.nihon-u.ac.jp
N i h o n U n i v e r s i t y
1-3-2 Misaki-Cho Chiyoda-Ku Tokyo 8 1 - 3 - 3 2 1 9 - 3 4 9 8

Summary

"The Dictionary of English Etymology for Natural Language Processing" (hereinafter referred to as DEE) contains the origin and borrowing process of approximately 20,000 English words and will be expected to expand to 250.00. The DEE has been developed together with a tagging utility program. The tagging program allows users to tag etymological labels to each word of the target text by consulting the DEE database. One object of the DEE is that it allows scholars automatically to analyze English expressions with the view that the etymology of each word in a text is one of the essential elements determining the behavior of the expression. Setting up etymology as one attribute of any word, researchers will be able to create the potential for another tool for semantic and rhetorical analysis.

Keywords Etymology, Etymological Labels, Tagging, Cognate Words, Rhetorical Collocations

Introduction

English words carry a variety of information such as spelling, pronunciation, meaning, word-class, and so on. These attributes have been subject to modification throughout the history of the English language. The etymological information, however, is somewhat different in that chronology of each word does not change throughout the history of English. Then, it is rather understandable when we posit that this unchangeable attribute has something to process a word behavior. In other words, the characteristics will shed new light upon the study of linguistic expressions in which the words in question are involved, when the analysis of the origin of words is properly made together with the semantic and rhetorical analysis.

Conventionally it has taken hours to tag the etymological labels to target texts, because researchers had to tag the labels manually [Tabata 1998]. DEE and tagging tools, however, will provide the many scholars with automatic labeling tool. We will show some experimental case studies illustrating the potentialities of the automatic etymology labeling.

1. The Dictionary of English Etymology for Natural Language Processing

The DEE is a database of the etymological information of English words. This database is originally designed to reduce the burden of looking up the origin of each word in the target text and then the origin of each word has been identified primarily by referring "*The Kenkyusha Dictionary of English Etymology* 1997", "*An Etymological Dictionary of English Derivatives* 1992"(Nihon Tosho Lib), "*The Oxford English Dictionary*, second edition", "*The Oxford Dictionary of English Etymology* 1966". The DEE contains the origin and the borrowing process of approximately twenty five thousand English. By consulting the DEE database, the tagging utility program tags etymological labels such as "**OE**" for Old English, "**OF**" for Old French, and "**L**" for Latin, to each word of the target text. As shown in ***Figure 1***, a spreadsheet software are employed as the database of the DEE. The sum total of words in the DEE is twenty five thousand six hundred and ten at present. However, not all the entries are accompanied by the etymological information yet. The headwords are listed in the column A, with their word-class in the column B. The column C, D, and E are concerned with the chronology of the headword. Thus, the column C shows the year when the headword was first used in the written material. It is possible to find out to what stage of English does the year in column C belong, when looking at the column D. The column E is prepared for the cases where the headword first appeared with a different word form, that is, spelling, from the one listed in the column A, but authors have not finished filling in the blanks yet, and temporarily put some other miscellaneous information in it. The actual etymology of the headword is registered from the column F onward. Thus,

the origin of the headword is shown in the column F, with its original word form at the time in the column G. We leave the cell blank if the spelling is identical to the one in the column A. When the item in this column is understood as a loan word, authors further trace the borrowing process of the word back to its true origin, and thus have column H and I, J and K, and so on.

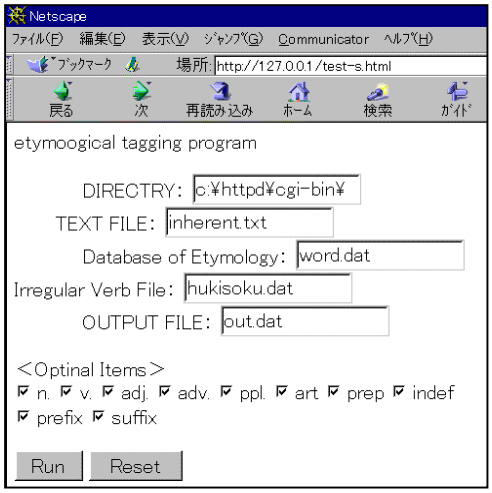
2. Tagging Utility Program

A tagging utility program have been also developed in order to take the full advantage of the DEE database. The program is a CGI file written in the Perl language, and it attaches etymological labels to words in target texts by consulting the DEE database. Since authors employ the HTML file format for the interface of the program, location of the target text files is done through the Web browsers as shown in ***Figure 2***. It is possible to direct to which word-class users want to attach the etymological labels. When the tagging program receives the "run" command, it generates an HTML file (and TEXT file) in which etymological labels are tagged to words, and then, outputs the result to the Web browser.

3. Experimental Applications of the DEE

The purpose of the DEE and its accompanying tagging utility is the automatic generation of the etymologically labeled text files so that the multidimensional analysis of English expressions will be easily attained. We will present some experimental examples in the analysis of English expressions with the help of the etymological labeling tool, herein.

Figure 2



3-1. Etymology and Rhetoric

The first experimentation is analysis of rhetorical expressions. We will find some etymological tendency, for example, in Antithesis. Rhetoric includes many techniques for expressing all matters relating to beauty or forcefulness of style. Writers prefer in general to invent rhetorical phrases such as *Antithesis, Hendiadys, Catalog, Intensifying Simile*, and so on. Tagging etymological labels to their rhetorical phrases, we will be able to know a high probability that in antithesis cognate words are paired as shown in ***Table 1***.

Figure 1

	A	B	C	D	E	F	G
1	WORD	WORD CLASS	Chronology			Back to	
2			YEAR	AGE	FORM	From	FORM
3	a	indef. art.		ME		OE	a, one
4	abac	adj.	1930	PE	abacus	L	abacus
5	abaciscus	Obs.		L		Gk	
6	abacist	adj.	1387	ME	abacista		
7	aback	adv.		ME		OE	
8	abaculus				abacicus		
9	abacus	n.	1387	ME		Gk	
10	abaft	adv., prep.	1275	ME		OE	
11	abalone	n.	1850	PE		Sp	abulon
12	abandon	v.	1375	ME		OF	abandoner
13	abandoned	ppl., adj.	1477	ME	< abandon	OF+ed	
14	abandonee	n.	1848	PE	abandonne		
15	abandonment	n.	1611	ModE	abandonnement	F	abandonnement
16	abase	v.	1477	ME		OF	abaissir
17	abasement	v.	1561	ModE		F	abaissement
18	abash	v.	1375	ME		OF	esbhier
19	abashment	n.	1410	ME		OF	abaissement
20	abasia	n.	1890	PE		mod L	abasie
21	abate	v.	1366	ME		OF	abatre
22	abate	v.	1860	PE		AF	abatre
23	abatement	n.	1528	ModE		OF	abatement
24	abatement	n.	1330	ME		AF	abatement
25	abattoir	n.	1820	PE		mod F	abattre

Table 1 Cognate: 44/50 Not Cognate(∗): 6/50

1.	alive(OE) or dead(OE)
2.	ancient(OF) and modern(OF)
3.	back(OE) and forth(OE)
4.	beginning(OE) and end(OE)
5.	birth(OE) and death(OE)
6.	black(OE) and white(OE)
7.	bride(OE) and groom(ME)
8.	∗cities(OF) and towns(OE)
9.	compassion(LL) and revulsion(L)
10.	∗both painful(OF) and blissful(OE)
11.	east(OE) and west(OE)
12.	friend(OE) or(OE) foe(OE)
13.	∗front(OF) and back(OE)
14.	front(OF) and rear(OF)
15.	good(OE) and bad(OE)
16.	good(OE) and evil(OE)
17.	heaven(OE) and earth(OE)
18.	head(OE) and tail(OE)
19.	heart(OE) and brain(OE)
20.	heaven(OE) and hell(OE)
21.	high(OE) and low(ON)
22.	husband(OE) and wife(OE)
23.	increase(OF) and decrease(OF)
24.	joys(OF) and disappointments(OF+of)
25.	∗joys(OF) and sorrows(OE)
26.	men(OE) and women(OE)
27.	mind(OE) and body(OE)
28.	life(OE) and death(OE)
29.	long(OE) and short(OE)
30.	∗losses(ME) and gains(OF)
31.	night(OE) and day(OE)
32.	north(OE) and south(OE)
33.	offensive(F) and defensive(OF)
34.	positive(OF) and negative(LL)
35.	pro(L) and con(L)
36.	reward(AF) and punishment(OF)
37.	rich(OF) and poor(OF)
38.	right(OE) and left(OE)
39.	rise(OE) and fall(OE)
40.	sooner(OE) or later(OE)
41.	spoken(OE) and written(OE) words(OE)
42.	∗supply(OF) and drain(OE) water(OE)
43.	sunrise(ME) and sunset(ME)
44.	theory(LL) and practice(OF)
45.	top(OE) and bottom(OE)
46.	top(OE) and tail(OE)
47.	ups(OE) and downs(OE)
48.	yeas(OE) and nays(ON)
49.	young(OE) and old(OE)
50.	work(OE) and play(OE)

AF:Anglo-French, Norw:Norwegian, ON:Old Norse, LL:Late Latin, ONF:Old Norman French, LG:Low German, MDu:Middle Dutch

Rhyme is the repetition of sou positions close enough to be noticed. Another aspect of rhyming means that the words in the rhetorical phrase have the same affixes and such matter indicates the words are cognate if comprising the same prefix or suffix. We can retrieve a lot of rhetorical collocations with conditional expressions of regular expression describing the same prefix or suffix, as follows:

- *connection and disconnection/creation and destruction*
- *addition and deletion*
- *addition, subtraction and multiplication*
- *stability, maneuverability and followability*

He becomes more callous, the *population* becomes more hostile, the *situation* grows more tense, and *police force* is increased. James Baldwin
population(LL)/*situation*(OF)/*police*(F) *force*(OF)

3-2. Etymology and Concepts

The second experimentation deals with the relationship between etymology and concept in Roget's Thesaurus. Peter Mark Roget proposed a methodology for compiling his thesaurus in the introduction to the original edition in 1852. 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 in ***Table 2***. The words have been tagged with the etymological labels. The table indicates German: 8 versus Romance: 32. This experimental result suggests that most of abstract vocabulary has the romance origin.

Table 2

Roget's Thesaurus Class I Abstract relations Section I.

#1. Existence. N. [German 8 : Romance 32]
[**Cluster 1** G 1 : R 5 (OE 1; OF 1, L- 4)] existence(OF), being(OE), entity(medL), ens(Lat), esse(Lat), subsistence(LL).
[**Cluster 2** G 1 : R 10 (OE 1; OF 7, L- 3)] reality(OF), actuality(medL); positiveness(OF+oe) &c. adj.; fact(L), matter(OF) of fact(L), sober(OF) reality(OF); truth(OE) &c. 494; actual(OF) existence(OF).
[**Cluster 3** G 0 : R 4(OF 4)] presence(OF) &c. (existence(OF) in space(OF)) 186; coexistence(OF+of) &c. 120.
[**Cluster 4** G 3 : R 3(OE 3; L 3)] stubborn(?OE) fact(L), hard(OE) fact(L); not a dream(OE) &c. 515; no joke(L)
[**Cluster 5** G 3 : R 9(OE 3; OF 5, L 1)] center(OF) of life(OE), essence(L), inmost(OE) nature(OF), inner(OE) reality(OF), vital(OF) principle(OF).
[**Cluster 6** G 0 : R 1(L- 1)] [Science of existence], ontology(newL).

"G" is Germanic origins, "R" is Romance origins; "+oe, +of" means the origin of the suffix; "L-" means the group of Latin origin; "Lat" indicates Latin words.

Table 2-a & 2-b shows "correlative words" in his "*Introduction*". On the triads in **Table 2-a**, Roget remarks "*two ideas which are completely opposed to each other, admit of an intermediate or neutral idea, equidistant from both*," and for those in **Table 2-b**, he says "*the same word has several correlative terms, according to the different relations in which it is considered*." The interesting point here is that each component of these triads or couples in most of the cases are of the same lineage. That is to say, there is a definite tendency that correlative words have the similar origin. This is the proof that words semantically related to each other assemble to the cluster similar in origin. It will be accordingly clear that expressions are not only under the government of syntactic and semantic properties, but also under the influence of their etymological background. In the light of this, we have been planning to tag the labels onto the whole of Roget's Thesaurus and then will be able to determine whether or not there is any etymological bias in each entry of the concept. We, for example, have found out the trend that Romance origin is prominent in Class I (words expressing abstract relations), while Germanic origin is overwhelming in Class VI (words relating to the sentiment and moral powers.)

3-3. Etymology and Style

The analysis of the origin of words has been commonly associated with stylistics. We have tried the automatic tagging of the etymological labels to the real text materials. Some examples are shown here, which include "*The End of Something*" by a writer Ernest Hemingway and "*The Modularity of Mind*" by a philosopher Jerry Fodor. It is widely known that most of mono-syllabic Germanic vocabulary is overwhelmingly used in Hemingway' works. The tagged text in ***Text 1*** confirms his Saxon-ism as shown clearly in ***Table 3***. To the contrary, in the tagged text of Fodor's, as an example of the scientific writing, the frequency of the Romance words outnumbers that of the Germanic as shown in ***Table 4***. Such contrast between these different types of lineage, as well as the contents, suggest that lyric expression includes many Saxon words and philosophical expression includes many Romance. Thus, etymological statistics provides the possibility of finding some regularity of writing style.

4. Another Application Perspective

When ***DEE*** and its utility can work in conjunction with text processing, 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*** will be simplify English complex sentences into simple sentences and replace words that are abstruse or abstract words (Romance origin) with ones that are plain or concrete (German origin). Thus, we will intend to contribute PSET (Practical Simplification of English Texts).

Concluding Remarks

The historical process of English language is reflected in its contemporary vocabulary and expressions. As many philologists pointed out [Ullman 1961, Brunner 1960, Scheler 1977], English language consists of three historical 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 related to the deepest in the mind of the native speaker and is tied to images of the soul [Peterson 1990, Stier1998], the ***French*** includes concepts of social consciousness and the ***Latin*** involves logos or reason. We are preparing a hypothetical scheme as depicted in ***Figure 3***. This conceptual diagram in this figure is based on the analogy with phylogeny and ontogeny in paleontology and will be able to enable to analyze English expressions from the etymological point of view and clarify the essential elements determining the behavior of the expression.

References

1. Tabata, T 1998 'The Generating Pattern of High Frequency Words in Dickens's Works' http://www.lang.osaka-u.ac.jp/~tabata/papers/1998/oct25.pdf
2. Ullman, S. 1961 "Semantics An Introduction to The Science of Meaning" Basil Blackwell & Mott Ltd.
3. Brunner, K.1960,1962 "Die englische Sprache: Ihre geschichtliche Entwicklung"
4. Scheler, M. 1977 "De englische Wortschatz" Berlin Erich Schmidt Verlag
5. Petersen, M. 1990 "Mark Remarks" Tokyo Iwanami
6. Stier, W. 1998 "Power Write Techniques of Writing in English" Tokyo Maruzen
7. Saraki 1999 "Building up The Dictionary of Rhetorical Collocations" pp. 399-401 Proc. of The 5th Annual Meeting of The Association for National Language Processing in Japan
8. Osada, Saraki 1999 'Text Analysis Using Etymological Database' Symposium held by SIG Language and Communication of The Institute of Electronics, Information and Communication Engineering in Japan http://www.etl.go.jp/etl/nl/sympo99/saraki1/paper.htm
9. Tait, J., Devlin, S *"PSET: Description of Proposed Research"* http://osiris.sunderland.ac.uk/~pset/pset-prop-shortened.html

Table 2-a *Correlative Words*

Identity(**L**)	Difference(**OF**)	Contrariety(**OF**)
Beginning(**OE**)	Middle(**OE**)	End(**OE**)
Convexity(**L**)	Flatness(**ON)**	Concavity(**L**)
Desire(**OF**)	Indifference(**L**)	Aversion(**L**)
Insufficiency(**lateL**)	Sufficiency(**L**)	Redundance(**L**)

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

Table 2-b *Correlative Words*

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**)
Acquisition(**medL**)	Deprivation(**medL**)/ Loss(**OE**)
Refusal(**OF**)	Offer(**OF**)/ Consent(**L**)
Use(**OF**)	Disuse(**OF**)/ Misuse(**OF**)
Teaching(**OE**)	Misteaching(OE?)/ Learning(**OE**)

Text 1 *"The End of Something"* Tagged

In the old(OE) days(OE) Hortons bay(OF) was a lumbering(ME) town(OE). No one(OE) who(OE) lived(OE) in it was out of sound(OE-ON) of the big(ME-ON) saws(OE) in the mill(OE) by the lake(OF). Then(OE) one(OE) year(OE) there(OE) were no more(OE) logs(ON?) to make(OE) lumber(ME). The lumber(OE) schooners(American) came(OE) into the bay(OF) and were loaded(OE) with the cut(OE) of the mill(OE) that stood(OE) stacked(ON) in the yard(OE). All(OE) the piles(OE) of lumber(OE) were carried(NF) away(OE). The big(ME) mill(OE) building(OE) had(OE) all(OE) it machinery(F) that was removable(OF+of) taken(OE) out and hoisted(MDu) on hoard(OE) one(OE) of the schooners(American) by the men(OE) who(OE) had(OE) worked(OE) in the mill(OE). The schooner(American) moved(AN) out of the bay(OF) toward the open(OE) lake(OF) carrying(NF) the two(OE) great(OE) saws(OE), the travelling(OF) carriage(F) that hurled(lowG) the logs(ON?) against the revolving(OF), circular(AF) saws(OE) ……

Text 2 *"The Modularity of Mind"* Tagged

FACULTY(OF) PSYCHOLOGY(modL) is getting(ON) to be respective(L+of) again after centuries(L) of hanging(OE) around with phrenologists(Gk+of/gk+of) and other dubious(L) types(lateL). By faculty(OF) psychology(modL) I mean(OE), roughly(OE+oe), the view(AN) that many fundamentally(modL+oe) different(OF) kinds(OE) of psychological(modL+of) mechanisms(modL) must(OE) be postulated(medL) in order(OF) to explain(L) the facts(L) of mental(L) life(OE).Faculty(OF) psychology(modL) takes(OE) seriously(OF+oe) the apparent(OF) heterogeneity(medL) of the mental(L) and is impressed(OF) by such prima facie(Latin) differences(OF) as between(OE), say(OE), sensation(F) and perception(OF), volition(F) and cognition(L), learning(OE) and remembering(OF), or language(OF) and thought(OE). Since, according(OF) to faculty(OF) psychologists(modL+of), the mental(L) causation(medL)……..

Table 3

Germanic origin	593	Romance Origin	111
OE	528	OF	81
ME	24	F	11
ModE	1	AF	7
ON	29	NF/ONF	4
LowG	1	AN	3
MLG	6	L	3
MDu	4	LateL	2

Table 4

Germanic	Romance
441	689

Figure 3