Studia Anglica Posnaniensia 42, 2006

LINGUISTICS

AMBIGUITY AND LANGUAGE EVOLUTION: EVOLUTION OF HOMOPHONES AND SYLLABLE NUMBER OF WORDS*

MIEKO OGURA – WILLIAM S-Y. WANG

Tsurumi University, Yokohama – Chinese University of Hong Kong
Project on Linguistic Analysis, University of California, Berkeley

ABSTRACT

We investigate the evolution of homophones and its relation to the evolution of syllable number of words, based on the quantitative analysis on the historical data and simulation. We suggest that homophones are the outcome of arranging form-meaning associations according to Zipf's law to maximize the referential power under effort for the speaker constraints. We also discuss the neural bases of ambiguity and the relation between ambiguity and robustness in language evolution.

Furthermore, we show that homophones are stable and cumulate with the times. To avoid creating homophones, syllable number of words increases, with more recent entry dates of words associated with more syllables. We also explore stability of homophones and instability of synonyms in children's acquisition process. The mechanism of the evolution of homophones and syllable length of words works cross-linguistically from the emergence of language and goes on at present.

1. Introductory remarks

1.1. Preliminaries

The existence and abundance of ambiguity in languages has intrigued linguists for a long time. If we view language as a coding system to encode meanings with signals, it would seem that language is not optimal, because in an ideal code one signal should correspond to exactly one meaning. When there are one-to-many correspondences between form and meaning, ambiguities arise. We assume that the number of meanings that humans can manipulate is infinite and there may exist a cognitive constraint on the number of forms that can be memorized. To meet semantic need, polysemy and homophony, which are major sources of ambiguity, are inevitable. In this study, we would like to investigate the evolution of homophony and its relation to the evolution of syllable number of words, based on the quantitative analysis on the historical data and simulation.

Anttila (1989: 184) states that all languages have homophony to different degrees, and one can never predict with complete confidence when a community or speaker will find it inconvenient enough to be corrected. Even when avoidance or correction of homophony does occur, there is no way of telling by what mechanism of change it will happen. Furthermore, there has been little study on the evolution of syllable number of words except the observation that word length is inversely related to the size of the phonological inventory (Nettle 1999: 144-147).

After defining polysemy and homophony, we will investigate degree of homophony based on the quantitative study in English and Japanese and simulation, and examine sources, word frequencies and lexical categories of homophones in section 2. In section 3 we will inquire into the mechanism of how homophones evolve and how syllable number of words increases to avoid creating homophones. In section 4 we will show that homophones are stable, based on Old and Middle English data. In section 5 we will explore the problem of why homophones are stable and synonyms are unstable in children's acquisition process. In section 6, we will give concluding remarks.

1.2. Definition of polysemy and homophony

The border between polysemy and homophony is not always clear. An important issue is the definition of a "lexical item" or "word". According to Leech (1974: 229-230), the lexical entry is defined as a trio of specifications morphological, syntactic, and semantic. Two useful ways to define "lexical item" are:

a) a bundle of lexical entries sharing the same morphological specification

b) a bundle of lexical entries sharing the same morphological specification and the same syntactic specification.

The second definition states that any two lexical entries related by conversion belong to different lexical items (e.g. *face* as noun and *face* as verb), while the first regards these as variants of the same lexical item. Conversion is the derivational process whereby an item is converted to a new word-class without the addition of an affix. This relationship may be seen as parallel to that between verb *acquit* and noun *acquittal* (Greenbaum – Leech – Svartvik 1972: 1009). Marchand (1969: 359-360) calls it zero-derivation. We may assume that a word created by conversion or zero-derivation belongs to a different lexical item from the base form in parallel with a derived item with an affix. Thus we adopt the second definition for lexical item proposed by Leech.

We would like to define polysemy as forms with many related meanings within the same lexical category, and homophony as two or more unrelated meanings getting the same form within the same lexical category, or two or more meanings getting the same form across the different lexical category. By this definition of homophony, lexical items related by conversion or zero-derivation are treated as homophones. Jespersen (1933: 73) calls such lexical items as “grammatical homophones”. Wordnet 2.0 (2003) defines polysemy within a lexical category.[1]

2. Homophones in Present-day English and Japanese

2.1. Degree of homophony

Based on the CELEX lexical database of English, version 2.5 (1995) and the LDC Japanese Lexicon (1997), we find that 11,980 or 22.8% of 52,447 types and 8,827 or 17.2% of 51,274 types[2] are homophones in Present-day English and Japanese respectively. We give the number of homophones classified according to the syllable or mora[3] number in columns and the number of words in a homophonous set in rows for English in Table 1a and for Japanese in Table 1b. We also give the total number of types classified according to the syllable or mora number for English in Table 2a and for Japanese in Table 2b. In English 4,743, or 70.2% of 6,761 one syllable words are homophones, and in Japanese 145, or 58.3% of 252 one mora words and 1,232, or 40.9% of 2,946 two mora words are homophones.

[1] We also give evidence to consider conversion is a source of homophones based on the discussion in section 3. Homophones are usually composed of 2 words. In CELEX database, 9482 or 79.1% of 11,980 homophones are composed of 2 words as shown in Table 1. But polysemy usually has more than 2 meanings. In WordNet database, polysemous nouns and verbs have 2.79 and 3.66 meanings on the average respectively. We may assume that in polysemous words more recent creation dates of meaning do not always associate with more syllables. Thus the behavior of homophones and polysemous words are different, and words created by conversion show the behavior of homophones.

[2] 29,419 proper nouns are excluded.

[3] Mora in Japanese is a unit that can be represented by one letter of *kana*. It is a unit for such syllables as $\left\{\begin{matrix}(\mathrm{w})\\(\mathrm{C})(\mathrm{j})\end{matrix}\right\}$V, syllabic nasal, and assimilated sound called *hatsuon*.

Table 1a. Number of homophones in English

	1 syl.	2 syls.	3 syls.	4 syls.	5 syls.	6 syls.	7 syls.	Total
2 words	3068	3888	1792	588	132	8	6	9482
3 words	900	477	123	60	3	0	0	1563
4 words	460	104	16	0	0	0	0	580
5 words	175	40	0	0	0	0	0	215
6 words	96	0	0	0	0	0	0	96
7 words	28	0	0	0	0	0	0	28
8 words	16	0	0	0	0	0	0	16
Total	4743	4509	1931	648	135	8	6	11980

Table 1b. Number of homophones in Japanese

	1 mora	2 morae	3 morae	4 morae	5 morae	6 morae	Total
2 words	26	446	1404	2268	108	12	4264
3 words	27	231	687	885	3	0	1833
4 words	32	132	420	520	0	0	1104
5 words	5	125	165	225	0	0	520
6 words	12	108	132	132	0	0	384
7 words	21	14	84	126	0	0	245
8 words	0	56	48	56	0	0	160
9 words	9	45	45	27	0	0	126
10 words	0	50	30	0	0	0	80
11 words	0	0	33	0	0	0	33
12 words	0	12	12	12	0	0	36
13 words	13	13	0	0	0	0	26
14 words	0	0	0	0	0	0	0
15 words	0	0	0	0	0	0	0
16 words	0	0	16	0	0	0	16
Total	145	1232	3076	4251	111	12	8827

Table 2a. Total number of types in English

Syllable number	Number of types
1	6761
2	18564
3	15195
4	7970
5	3000
6	711
7	188
8	37
9	13
10	5
11	2
12	1
Total	52447

Table 2b. Total number of types in Japanese

Mora number	Number of types
1	252
2	2946
3	11108
4	18456
5	8343
6	5204
7	2567
8	1344
9	500
10	275
11	128
12	76
13	37
14	24
15	9
16	3
17	1
23	1
Total	51274

Wang *et al.* (2004) report some modeling for homophony. The simulation model is designed within the naming games framework proposed by Steels (1996). Agents in the model are assumed to be able to produce a number of distinctive utterances and to make use of such utterances to communicate a set of meanings. Agents can create new words at random, as well as learn the words created by other agents. An example of two matrices with three meanings (m1, m2, m3) and three utterances (u1, u2, u3) is given in Table 3. Each element of the matrices represents the probability that an agent has an association between a certain meaning and a certain utterance. The two matrices have the constraint that each row of the speaking matrix and each column of the listening matrix sum to one, to meet the assumption that each meaning is expressible, and each utterance is interpretable. At the beginning speaking and listening matrices of each agent are randomly initialized.

Table 3. An example of speaking and listening matrices in the interaction model (Ke *et al.* 2002)

	u1	u2	u3		u1	u2	u3
m1	0.3	0.4	0.3	m1	0.1	0.3	0.6
m2	0.4	0.55	0.05	m2	0.5	0.3	0.3
m3	0.7	0.2	0.1	m3	0.4	0.4	0.1
	Speaking matrix				Listening matrix		

At each step, two agents are chosen to communicate, one as the speaker and the other as the listener. The speaker decides a meaning he wants to communicate, looks for or creates an utterance which is associated with the meaning, and transmits the utterance to the listener. The listener perceives the utterance and tries to interpret the meaning by searching his existing vocabulary. If he interprets the same meaning for the utterance, then this is considered to be a successful communication. Each word has a score; after each successful communication, the score of the word is increased. Otherwise, the score is decreased. When the score of the word becomes too small, the word is removed from the vocabulary. Upon failure, the listener learns the word from the speaker by adding an association between the perceived utterance and the intended meaning of the speaker. After a number of interactions, we observe that associations between objects and utterances are shared by all agents.

Figure 1 shows simulation results when the number of meanings (M) and the number of utterances (U) are equal. In this simulation only one word, i.e. a word without any context, is transmitted during the communication. There are 10 agents, initialized randomly, with a vocabulary size of M=U=5. We can see that agents are able to acquire the same vocabulary, and their communications are successful 90% of the time, 20% of the words having homophones. When we compare the quantitative studies stated above with the simulation results, we may assume that to avoid homophones, humans try to manifest “one meaning, one form”, but homophones do occur and the threshold is around 20% of the vocabulary.[4]

[4] Ke *et al.* (2002) also design an imitation model to simulate the emergence of a shared vocabulary. In this model, there are a number of agents, each of which initially has its own set of mappings between meanings and utterances. When two agents interact, one imitates the other according to some strategy, either by random or by following the majority in the population, and the agents converge to an identical vocabulary. The following table shows the average probability of the vocabulary with one-form-one-meaning by Strategy 1 (S1), i.e. random imitation and by Strategy 2 (S2), i.e. imitation by following the majority, when the number of meanings is fixed: M=10, and population size (P) and number of utterances (U) are varied. For example, when U=10, i.e. M=U, and P=10 by S1, the probability of the vocabulary with one-form-one-meaning is 0.77, i.e. the probability of homophones is 0.23. We find that the probability of homophones is

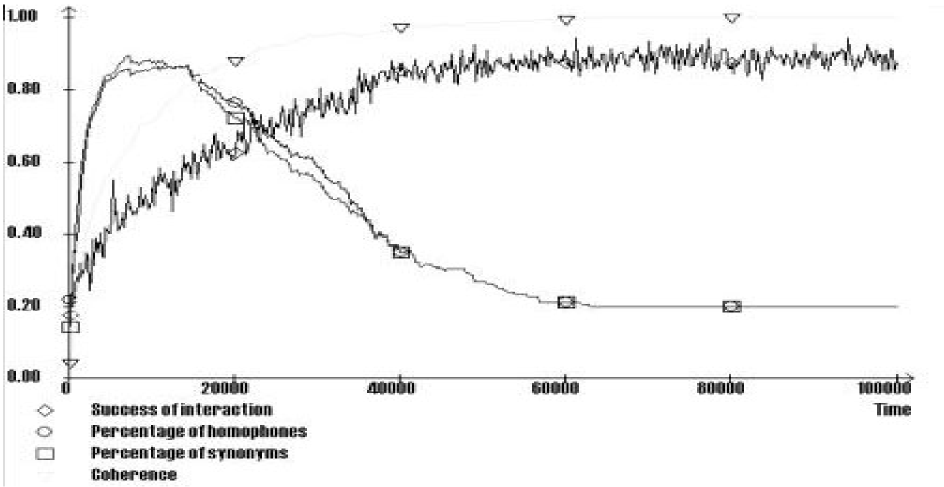

Figure 1. Simulation of homophones when M=U (Wang *et al*. 2004)

There are subtypes of iconicity, isomorphism, which denotes the one-form-one-meaning condition, and automorphism, which holds that linguistic elements which are alike semantically should also resemble one another formally (Haiman 1985: 4). The former type of iconicity underlies homophony, while the latter type underlies polysemy. Because of automorphism of iconicity, it is quite plausible that there are many polysemous words. But as for homophony, people try to avoid creating homophones because of isomorphism of iconicity.

Why do homophones occur even though humans try to manifest one-to-one correspondence between form and meaning? Why is the threshold of homophones around 20% of the vocabulary? Zipf (1949) suggests the simultaneous

around 0.25-0.20 when M=U, regardless of the population size, and even when the number of utterances outweighs the number of meaning by five times homophones seem to be unavoidable. It is interesting that the probability of homophones is almost the same whether we resort to the interactive model or the imitation model.

Table 13. Average probability of the vocabulary with one-form-one-meaning by Strategy 1 and Strategy 2 (adapted from Ke *et al*. 2002)

	S1			S2		
	U=10	U=30	U=50	U=10	U=30	U=50
P=10	0.77	0.91	0.94	0.74	0.92	0.95
P=30	0.76	0.92	0.95	0.73	0.92	0.95
P=50	0.78	0.92	0.95	0.74	0.91	0.94

minimization of the two opposing forces from listener and speaker for form and meaning associations. One form for all meanings is an ideal code for the speaker, while one-to-one correspondence between form and meaning is an ideal code for the listener. He supposes that Zipf's law, which states that word frequencies decay as a power law of its rank (see section 2.2.), is the outcome of form-meaning associations adopted for complying with listener and speaker needs. Arranging signals according to Zipf's law is the optimal solution for maximizing the referential power under effort for the speaker constraint. Thus we may assume that humans try to manifest one meaning, one form to avoid creating homophones, but at the same time they try to maximize referential power under effort for the speaker constraint, using a portion of words frequently and even forming homophones with high frequency words.

Ferrer i Cancho – Sole (2003) show that Zipf's law is a hallmark of referentially random systems and symbolic reference systems by using mathematical modeling. In our simulation of the emergence of a shared vocabulary, the same set of associations between meanings and utterances are formed from random creation by each agent. We may assume that the emergence of a shared vocabulary is the result of Zipf's law, which is found in the transition between referentially random systems created randomly by each agent and symbolic systems of a shared vocabulary, and the threshold of the homophones is around 20% of vocabulary, which is the limit of high frequency words (see Table 5 and Figure 2 in section 2.2.).

Our daily communication is not much hampered by homophones, because we generally do not process sentences in isolation from contexts. One may criticize our quantitative work, arguing that instances that show homophonic clash, thus cause ambiguity must be examined on a case-by-case basis.[5] However, as shown in Tables 1a and 1b, together with Tables 2a and 2b, the number of homophones decreases as the number of words in a homophonous set increases, and the percentage of homophones for each syllable or mora number decreases as the syllable or mora number increases. The distributions of homophones in Tables 1a and 1b suggest a threshold or limit of homophones that can be tolerated.[6] They also suggest that there is greater probability of ambiguity as the number of words in a homophonous set and the syllable or mora number increase.

We consider that all the different meanings of homophones must have been

[5] Kempson (1980) analyzes sentence ambiguities in homonymy and polysemy in the interaction between context and interpretation of the lexical item.

[6] Strang (1980) collects some 1700 monosyllabic words from the *OED*, and finds that the level of exploitation of homophones is higher than might be thought tolerable in several forms: /bi:/, /bɛ:/, /bei/, /bʌt/, /bʌk/. For example, in /bʌt/, there are 21 homophones, including 16 nouns and 4 verbs, 12 surviving till the 19th century.

largely analyzed before the listener is able to select the proper one according to the context. The brain puts all sorts of competing pieces of information in subconscious temporary storage until the context allows the appropriate one to surface. Thus, if there are many competing meanings in the subconscious temporary storage in the brain, the context no longer allows the appropriate one to surface, and ambiguity arises. Ongoing work by neuroscientists, e.g. Robert Desimone gives evidence to this effect from visual object recognition.

Robustness is the ability of the signal to withstand noise. Given an ambiguous utterance with several meanings, the meaning intended by the speaker is the signal and the other possible meanings are noise. Everything else being equal, a language with fewer ambiguous utterances is more robust – it has a higher signal to noise ratio. The avoidance of homophony contributes to robustness of language. Lass (1980: 75-80, 1987) argues against homophonic clash or avoidance of homophony as an explanation for sporadic changes. Our work is based on the quantitative examination of interaction of homophones and syllable length, which leads to growth in complexity in languages over time (see section 3). Speakers avoid homophones by taking avoiding strategy in advance and the development can be predicted. These are what Lass claims for the explanation by homophonic clash or avoidance of homophony.

As basis of the analysis of the present-day material, there is no practical alternative to Standard English. Dialectal items have wondered into and out of the lexical standard, and there is no record that makes clear, concerning every word that has declined into dialect, just when the loss of currency in standard occurred. Moreover, there was no variety in Old and Middle English corresponding to present-day Standard English. However, whichever variety we are concerned with, we assume that the same fundamental mechanism of interaction of homophones and syllable length works in it.

2.2. Sources, word frequencies, and lexical categories of homophones

Creation of homophones might be the effect of (1) conversion from one lexical category to another, (2) borrowing, (3) sound change and (4) innovation of new words. Table 4 shows the number of occurrence of words for each source in English homophones, classified according to date of creation. We also give percentage of occurrence for each source in each period. The data is based on the first 4,919 samples out of 11,980 homophones from the CELEX database. All the historical information is based on the *Oxford English dictionary*, *Version 2.0 on CD-ROM* (*OED2*). The original words in 2,260 homophonous sets of the 4,919 samples are excluded in calculation, thus the number of source words for homophones is 2,659 in total. The number in parentheses for conversion is the number of occurrence of words where conversion from compounds, onomatopoetic words, exclamatory words is excluded. We find that most of the homophones in Old English (OE) (c.700-1100) result from conversion, and those after the 12th century result from conversion and borrowing. Based on the number of occurrence of homophones created by conversion and borrowing, we find that there are two turning points: an upward one in the 16th century and a downward one in the 20th century. This agrees with the Strang's finding (1980) in monosyllabic words created by zero-derivation.

Table 4. Sources of English homophones

	Conversion		Borrowing		Sound change		Innovation	
	Number	%	Number	%	Number	%	Number	%
OE	171 (167)	90.48	6	3.17	2	1.06	10	5.29
12th c.	20 (18)	66.67	6	20.00	4	13.33	0	0.00
13th c.	83(76)	54.96	67	44.37	1	0.66	0	0.00
14th c.	152(140)	50.00	149	49.01	2	0.66	1	0.33
15th c.	133(118)	64.56	62	30.10	11	5.34	0	0.00
16th c.	279(206)	56.71	183	37.19	2	0.41	28	5.69
17th c.	213(163)	52.99	175	43.53	0	0.00	14	3.48
18th c.	122(76)	53.51	95	41.66	2	0.88	9	3.95
19th c.	271(144)	58.91	167	36.30	9	1.96	13	2.83
20th c.	137(44)	69.54	59	29.95	0	0.00	1	0.51
Total	1581(1152)	59.46	969	36.44	33	1.24	76	2.86

Some examples of homophones composed of original words and source words are given in (1). They are arranged with date of entry, lexical category, and source: (1) conversion, (2) borrowing, (3) sound change, (4) innovation of new words. Original words are shown without sources. When date of creation of homophone by sound change is different from date of entry of the word, it is given in parentheses. OE forms are given in parentheses for the homophones in (1c).

1a)	*beat*	885 (1400)	V	3
	beet	1000	N	
	beat	1400	Adj	3
	beat	1615	N	1 (< V)

1b)	*great*	888	A	
	great	950	N	1 (< A)
	grate	1400 (16c.)	N	3
	grate	14.. (16c.)	V	2 (< French), 3

1c)	*ewe* (*eowu*)	700	N	
	yew (*eow*)	725	N	4
	you (*eow*)	897	Pron	4
1d)	*encounter*	1297	N	(< French *encontre*)
	encounter	1300	V	2 (< French *encontrer*)
1e)	*curve*	1594	V	(< Latin *curvare*)
	curve	1696	N	2 (< Latin *curvus*)

Jespersen (1933: 73) states that conversion from one lexical category to another is one of the most characteristic traits of English, and is found to a similar extent in no other European language. Many cases of conversion are from N to V or from V to N. There are 547 samples, or 47.5% where V is converted from N, and 292 samples, or 25.3% where N is converted from V out of 1,152 samples for conversion (compounds, onomatopoeic words, exclamatory words are excluded). Many cases of borrowing are from French or Latin. There are 504 samples, or 52% from French, 281 samples, or 29% from Latin out of 969 samples for borrowing. In homophones resulted from borrowing, the original word is also a borrowed word in many cases. When the words were borrowed into English, simplification of the inflectional endings occurred and they became homophones. We decide whether homophones composed of borrowed words are created by borrowing or conversion from the original word, based on the description in the *OED2*.

The average word frequency for 2,260 original words in homophonous sets is 1,624, and that for 2,659 source words is 1,085. Table 5 gives the average word frequencies ranked from most frequent to least frequent in the CELEX database,[7] and Figure 2 plots the power-law distribution of word frequencies as a function of word rank on log-log coordinates. This graph shows a similar slope to Zipf's law that gives a straight line with a slope of -1. The average frequencies for original and source words of homophones are in word rank of 1001-2000 and 2001-3000 respectively. Thus we may state that the words that compose the homophones are high frequency words.

[7] The CELEX frequencies are based on the 17.9 million token COBUILD/Birmingham corpus.

Table 5. Average word frequencies ranked from most frequent to least frequent in the CELEX database

Word rank	Average frequency
1-500	26768.54
501-1000	2627.33
1001-2000	1262.48
2001-3000	674.13
3001-4000	430.35
4001-5000	294.31
5001-10000	140.55
10001-15000	53.26
15001-20000	26.24
20001-25000	13.57
25001-30000	6.77
30001-35000	2.97
35001-38731	1.13
38732-52447	0.00

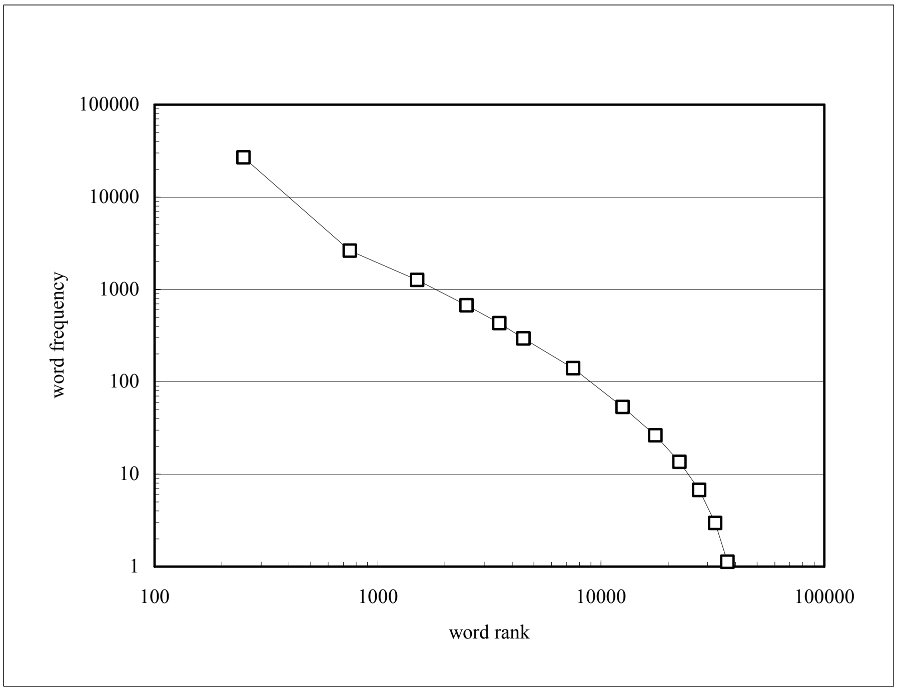

Figure 2. Power-law distribution of word frequencies in the CELEX database

We would like to add that most of the homophones are composed of words from different lexical categories in English. The number of occurrence of words whose lexical categories occur once in each homophonous set is 4,263 or 86.7% of 4,919 homophones. In Japanese, however, most of the homophones are composed of words from the same lexical categories as shown in (2). The number of occurrence of words whose lexical categories occur once in each homophonous set is 365 or 4.1% of 8,902 homophones. Unfortunately there is no such dictionary as *OED* in Japanese and we cannot trace the sources of homophones.

2a)	*akusei*	'malignancy'	N
	akusei	'misgovernment'	N

2b)	*akeru*	'to dawn'	V
	akeru	'to open (vt)'	V

3. Evolution of homophones and syllable number of words

3.1. Evolution of homophones

The data are based on the first 4,919 samples out of 11,980 homophones from the CELEX database of English, and historical information is based on the *OED2*. Table 6 shows the number of occurrence of homophones that arose from OE to the 20th century. The total number of occurrence of homophones is classified according to the syllable number, and the mean syllable number of the homophones is given in each period. Figure 3 shows the cumulative number of homophones from OE to the 20th century. The abscissa shows date of creation of homophones, and the ordinate shows the cumulative number of homophones. Table 6 and Figure 3 are based on the homophones that exist in Present-day English, and it is not clear how many homophones became obsolete. However, we find, based on the study on the evolution of homophones in OE and Middle English (ME) (c1100-1400) in section 4, that homophones are stable and cumulate with the times. Thus we assume that the number of occurrence of homophones in each period reflects the number of homophones that were created in that period.

Table 6. Number of occurrence of homophones that arose from OE to the 20th century

Date	Number of occurrence								Mean syl. num.
	Total	1 syl.	2 syls.	3 syls.	4 syls.	5 syls.	6 syls.	7 syls.	
OE	674	484	174	15	1				1.30
12th c.	80	46	29	5					1.49
13th c.	397	204	155	37	1				1.58
14th c.	681	268	290	96	27				1.83
15th c.	359	161	138	43	16	1			1.77
16th c.	815	294	314	156	43	8			1.97
17th c.	630	173	232	162	47	16			2.21
18th c.	340	84	134	88	20	11	2	1	2.26
19th c.	667	165	229	183	71	18	1		2.33
20th c.	276	49	104	93	20	8	1	1	2.42

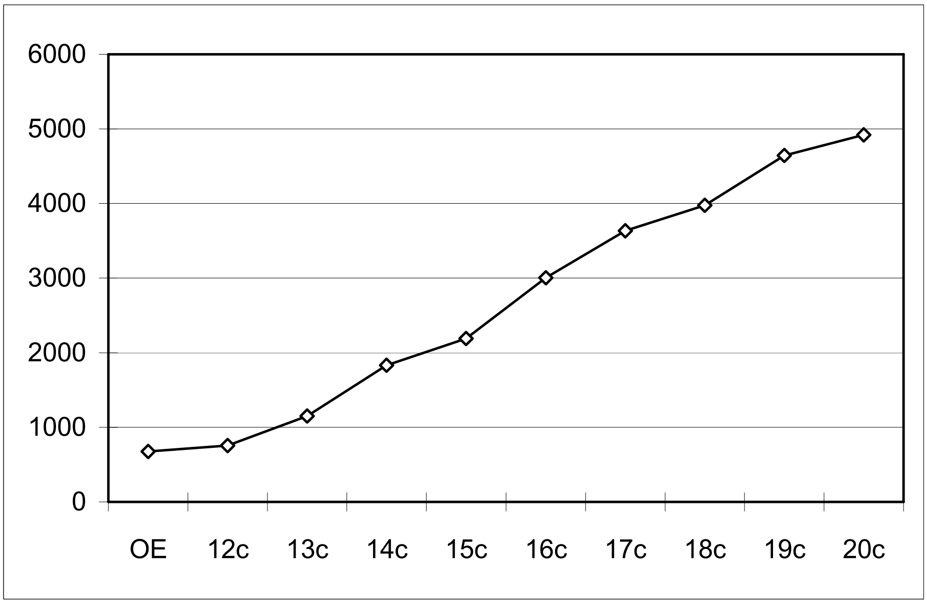

Figure 3. Cumulative number of homophones from OE to the 20th century

3.2. Evolution of syllable number of words

Homophones cumulate with the times, and there is a correlation between syllable number and date of creation of homophones, with more recent creation dates associated with more syllables as shown in the mean syllable number in Table

6. Then, how are the homophones avoided? We would like to show that syllable number of word increases to avoid creating homophones. Figure 4 plots mean syllable number of nouns and verbs with mean syllable number of homophones as a function of their date of entry into English. There are 2,874 nouns and 853 verbs in our database. We obtained them from 5,245 samples, which are every eleven samples of 52,447 types in the CELEX database. Each word was classified according to its date of entry, which was checked by the *OED2*. Table 7a shows number of occurrence of 2,874 nouns and 853 verbs, which are classified according to syllable number and their date of entry. Table 7b gives the mean syllable number of nouns and verbs in each period. From Figure 4 we find a correlation between syllable number and date of entry for both nouns and verbs, with younger words containing more syllables. We assume that syllable number increases, with more recent entry dates of words associated with more syllables, to avoid creating homophones.[8]

[8] The mean syllable number of homophones that were created in Old English (OE) might have been a little longer than that given in Table 6, which is based on Present-day English. Post-tonic vowels was in many cases pronounced with /ə/ in the 11th century, which ceased to be pronounced in the course of the Middle English period. Thus some homophones that were created in OE were one syllable longer in OE. This discussion also applies to the syllable length of nouns and verbs that entered in OE. However, the important fact is that the syllable number of nouns and verbs that entered in OE was larger than that of homophones.

One might argue that many nouns and verbs that entered after the 12th century are the borrowing from Graeco-Romance lexicon, which would have increased syllable number since there is virtually no simplex Germanic lexis with more than two syllables. However, as shown in Table 4, a large number of homophones also result from Graeco-Romance lexis after the 13th century. Some were created by the borrowings that entered in the preceding period. Thus the syllable number of homophones increases with the times, and so does the syllable number of nouns and verbs to avoid creating homophones.

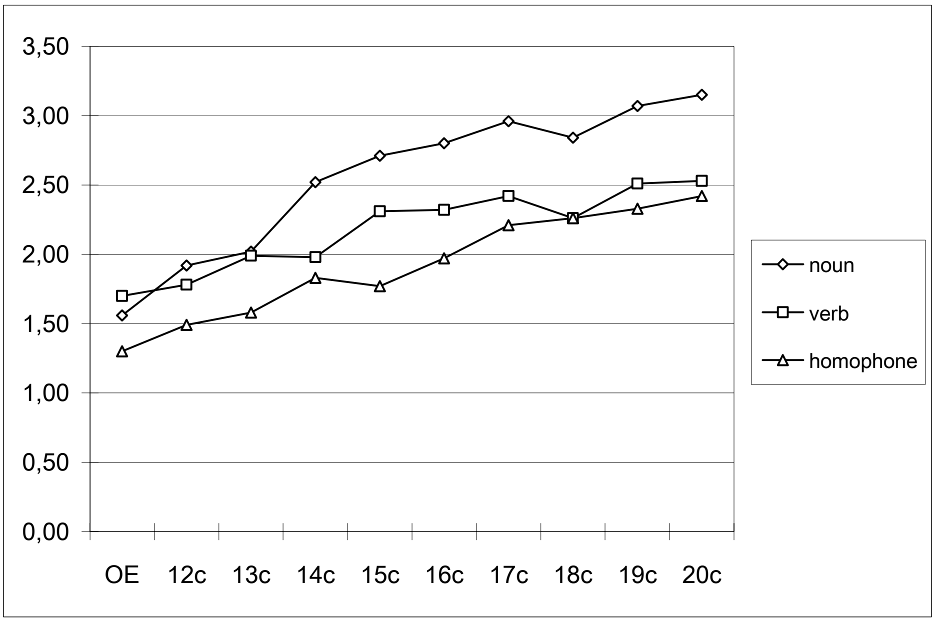

Figure 4. Mean syllable number of nouns, verbs and homophones as a function of their date of entry

Table 7a. Number of occurrence of nouns and verbs classified according to syllable number and date of entry: Nouns

Date	1 syl.	2 syls.	3 syls.	4 syls.	5 syls.	6 syls.	7 syls.	8 syls.	9 syls.	10 syls.	11 syls.	Total
OE	95	59	15	2								171
12th c.	28	28	12	3		1						72
13th c.	34	62	23	5	1							125
14th c.	39	142	100	40	10							331
15th c.	19	68	70	32	7	1						197
16th c.	27	129	125	56	20	5						362
17th c.	28	111	116	74	31	6						366
18th c.	19	115	89	40	16	9	2					290
19th c.	17	222	196	126	47	17	7					632
20th c.	11	118	102	57	15	12	6	3	3		1	328
Total	317	1054	848	435	147	51	15	3	3	0	1	2874

Table 7a. occurrence of nouns and verbs classified according to syllable number and date of entry: Verbs

Date	1 syl.	2 syls.	3 syls.	4 syls.	5 syls.	Total
OE	59	56	15	2		132
12th c.	18	21	4	2		45
13th c.	17	52	14	1		84
14th c.	45	85	31	5		166
15th c.	10	40	20	7		77
16th c.	26	43	34	13	1	117
17th c.	13	43	32	10	1	99
18th c.	7	17	12	1	1	38
19th c.	8	19	26	6		59
20th c.	2	19	10	4	1	36
Total	205	395	198	51	4	853

Table 7b. Mean syllable number of nouns and verbs

Date	Noun	Verb
OE	1.56	1.70
12th c.	1.92	1.78
13th c.	2.02	1.99
14th c.	2.52	1.98
15th c.	2.71	2.31
16th c.	2.80	2.32
17th c.	2.96	2.42
18th c.	2.84	2.26
19th c.	3.07	2.51
20th c.	3.15	2.53
Average	2.75	2.13

We also find that mean syllable number of nouns is larger than that of verbs after the 14th century, while they are almost the same before the 13th century. This is due to much larger number of occurrence of nouns and their distribution from 1 syllable to much larger syllable number than verbs after the 14th century. From the distribution of occurrence of nouns and verbs at each syllable number in each century shown in Table 7a, we find that the greater the total number of occurrence of nouns than verbs in each century, the greater the range of distribution of the syllable number. The average ratio of the total number of occurrence of nouns and verbs is 1.4 (368 nouns/261 verbs) before the 13th century, and 4.2 (2,506 nouns/592 verbs) after the 14th century. Thus the mean syllable number of nouns is larger than that of verbs after the 14th century.

The following are total number of occurrence and mean syllable number of nouns and verbs in open class words and conjunctions and pronouns in closed class words, based on the 38,920 samples from the CELEX database where parts of speech are given.

	total number of occurrence	syllable number
noun	22,205	2.53
verb	5,441	1.93
conjunction	50	1.78
pronoun	133	1.74

Number of occurrence of closed class words is much smaller than that of open class words, and within open class words, number of occurrence of nouns is 4.1 times as large as that of verbs. Syllable length of closed class words is shorter than that of open class words, and within open class words, nouns are longer than verbs. There is a correlation between syllable number and total number of occurrence of words. Most of the closed class words entered in OE and ME and they remain short in Present-day English.

The above observation is confirmed in the Japanese data. The following are the total number of occurrence and the mean mora number of nouns and verbs in open class words and conjunctions and pronouns in closed class words, based on the 51,274 samples of the LDC Japanese Lexicon.

	total number of occurrence	mora number
noun	40,846	4.44
verb	5,441	4.11
conjunction	87	3.24
pronoun	98	3.49

There is no consonant cluster in Japanese, thus the size of phonological inventory is smaller in Japanese than English. Word length is inversely related to the size of the phonological inventory, thus mora length of words in Japanese is longer as a whole than syllable length in English.

4. Homophones in Old English and Middle English

4.1. Stability of OE homophones

The *Brooklyn – Geneva – Amsterdam – Helsinki parsed corpus of Old English* (2000) contains 18,629 types, and 1205 types, or 6.5% of them are homophones. Scoring criterion is different from the CELEX database of English and the LDC Japanese Lexicon. Any differences in forms are counted as different types. For example, homophones *al* as adjective and pronoun are treated as different from homophones *all* as adjective and pronoun, both of which remain as “all” in Present-day English. Also different inflectional forms are counted as different types. For example, *sum*, which remains as “some” in Present-day English, has the following 8 homophonous sets: *sum*, *suman*, *sume*, *sumere*, *sumne*, *sumon*, *sumre*, *sumum* as adjective and pronoun. They are treated as one homophonous set “some” as adjective and pronoun in the CELEX database.

Some examples of OE homophones are given in (3). Words with @ exist in Present-day English, those with * are archaic, literary or dialectal and those with + are obsolete. Words within quotation marks are those that are used in Present-day English.

3a)	*ægðer*	‘either’	Conj.	@
	ægðer	‘either’	Pron.	@
	ægðer	‘either’	Adj.	@
3b)	*sæde*	‘say’	V	@
	sæde	‘seed’	N	@
3c)	*stille*	‘still’	Adj.	@
	stille	‘still’	Adv.	@
3d)	*an*	‘one’	Adj.	@
	an	‘one’	Pron.	@
	an	‘one’	Prep.	+
	an	‘one’	Adv.	+
3e)	*halga*	‘holy’	Adj.	@
	halga	‘hallow’	N*	

780 types or 65% of 1,205 OE homophones are still used in Present-day English, and 425 types became obsolete. Table 8 gives the number of types that became obsolete in each period. The dates of obsolescence are checked by the *OED2*. Many types of the 144 homophones that became obsolete in the 19th century are archaic, literary and dialectal in Present-day English. We also find that in 487 or 85% of 572 homophonous sets, at least one type exists in Present-day English. Thus we may state that OE homophones are stable. Many exist in Present-day English and some persist long.

Table 8. Number of occurrence of OE homophones that became obsolete

OE	37
12th c.	13
13th c.	68
14th c.	43
15th c.	49
16th c.	23
17th c.	35
18th c.	13
19th c.	144

4.2. Stability of ME homophones

We would like to confirm the above observation on the OE homophones in ME homophones based on the *Penn – Helsinki parsed corpus of Middle English* (2000). This corpus contains 48,725 types,[9] and we find that 4,653 or 9.6% of them are homophones.

2,691 types (= 1,966 types of OE origin + 725 types of ME origin) or 57.8% of 4,653 ME homophones (= 2,981 types of OE origin+1,672 types of ME origin) are used in Present-day English. Table 9 gives the number of types that became obsolete from the 12th to the 19th century, classified according to OE origin and ME origin. The total of the types of OE origin that became obsolete in the 19th century and those that are still used in Present-day English are 2,669 types (703 types + 1966 types) or 89.5% of 2,981 homophones of OE origin. The total of the types of ME origin that became obsolete in the 19th century and those that are still used in Present-day English are 1,377 types (652 types + 725 types) or 82.4% of 1,672 homophones of ME origin. It is not clear how many homophones of OE origin became obsolete in OE from Table 9, but we may confirm that homophones of both OE and ME origin are stable, and homophones of ME origin cumulate on those of OE origin.

[9] 2003 clitics are excluded.

Table 9. Number of occurrence of ME homophones that became obsolete

Date	OE origin	ME origin
12th c.	8	7
13th c.	61	29
14th c.	48	33
15th c.	73	49
16th c.	57	48
17th c.	52	92
18th c.	13	37
19th c.	703	652

4.3. Word frequencies of OE and ME homophones

Tables 10a and 10b give the average word frequencies ranked from most frequent to least frequent in the OE and ME database respectively. Figure 5 plots the power-law distribution of word frequencies in OE and ME database on log-log coordinates. The slopes of graphs are not so steep as -1 shown in Zipf's law. This is because word frequencies of OE and ME database are based on scoring criterion of types where any differences in forms are counted as different types (see section 4.1.). The average frequencies of OE and ME homophones are 30.7 and 87.9 respectively, and they are high frequency words in word rank of 251-500 of OE database in Table 10a and word rank of 501-1000 of ME database in Table 10b respectively.

Table 10a. Average word frequencies ranked from most frequent to least frequent in OE database

Word rank	Average frequency
1-250	216.53
251-500	29.09
501-1000	15.39
1001-1500	9.29
1501-2000	6.62
2001-3000	4.61
3001-4000	3.24
4001-5000	2.45
5001-10000	1.43
10001-18629	1.00

Table 10b. Average word frequencies ranked from most frequent to least frequent in ME database

Word rank	Average frequency
1-250	1520.26
251-500	164.99
501-1000	78.76
1001-1500	43.21
1501-2000	29.52
2001-3000	19.80
3001-4000	13.30
4001-5000	9.88
5001-10000	5.62
10001-15000	2.89
15001-20000	2.00
20001-25000	1.17
25001-48725	1.00

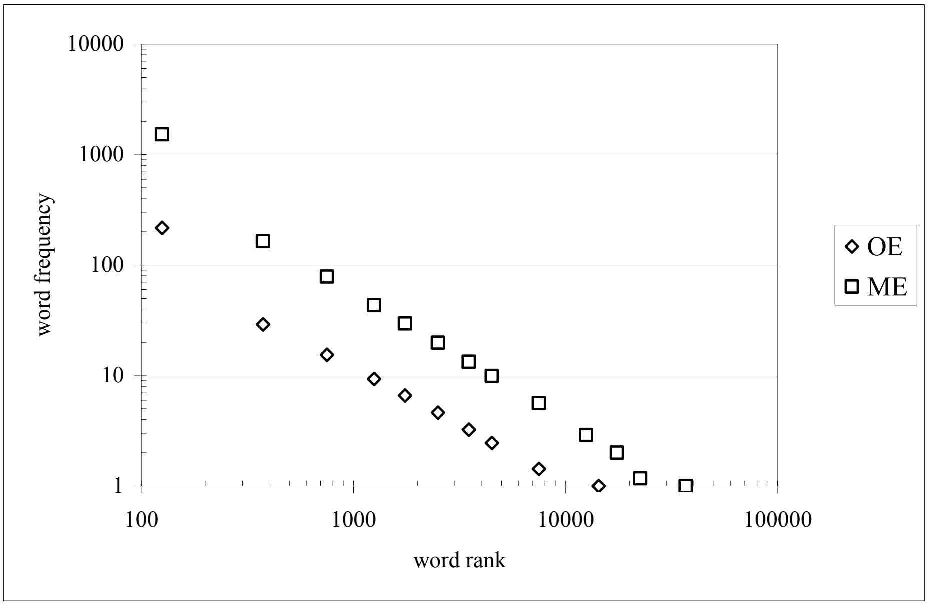

Figure 5. Power-law distribution of word frequencies in OE and ME database

Table 11 gives the average frequencies of the obsolete words from OE to the 19th century in OE database. We also give the average frequency of the words that remain in the 20th century. Table 12 gives the average frequencies for the obsolete words from 12th to the 19th century in ME database, classified according to OE origin and ME origin. We also give the average frequency for the words that remain in the 20th century. We find that words that became obsolete in OE in OE database in Table 11 are low frequency words. Word frequencies change with the times. From ME database in Table 12, we find that there is the correlation between dates of obsolescence and word frequencies. Words that became obsolete in the 19th century and those that remain in the 20th century are by far the most frequent words both for OE and ME origin. It is noted that the average frequency of words that became obsolete in the 13th century in OE database is high as compared with that in ME database. This may be due to the change that the relative pronouns, *þe* and *se*, which were frequently used in OE, were replaced by *þat* in the 13th century.

Table 11. Average frequencies of obsolete words from OE to the 19th century and of words that remain in the 20th century in OE database

OE	2.27
12th c.	29.08
13th c.	89.06
14th c.	45.58
15th c.	41.04
16th c.	10.65
17th c.	11.23
18th c.	39.77
19th c.	19.45
20th c.	28.98

Table 12. Average frequencies of obsolete words from the 12th to the 19th century and of words that remain in the 20th century in ME database

	OE origin	ME origin
12th c.	17.63	5.71
13th c.	9.10	3.14
14th c.	5.00	3.79
15th c.	47.84	19.27
16th c.	22.16	3.42
17th c.	21.58	3.40
18th c.	5.92	2.97
19th c.	66.15	33.55
20th c.	160.56	22.63

5. Children's acquisition of homophones and synonyms

We have shown that the homophones are stable in sections 3 and 4. However, synonyms are unstable. Ogura (1999), based on *A thesaurus of Old English* (*TOE*) (1995), shows that many words within near-synonyms in OE became obsolete in OE or ME.

Nowak *et al*. (1999) explains the stability of homonyms and instability of synonyms by the children's leaning mechanism. Individuals acquire a language by observing and imitating other individuals. Each individual undergoes a learning phase during which it constructs an association matrix, *A*. *A*'s entries, a_{ij} specify how often an individual has observed one or several other individuals referring to object *i* by producing signal *j*. The matrix *P* contains the entries P_{ij}, which denote the probability that for a speaker object *i* is associated with sound *j*, and are derived from the association matrix by normalizing rows and columns.

Homonymy refers to the case where there are more than one entries with 1 in a column, which shows a signal, of the matrix P_0 of the language, thus two different objects, i.e. meanings, which are shown in rows, are associated with the same word as shown in (4). We assume that when offspring receive k samples for each meaning from parent, the P_0, speaking matrix of the homophone leads to the *A*(ssociation) matrix that the child constructs during the learning phase as shown in (5), which again leads to the same P_1 matrix as shown in (6).

$$(4)\quad P_0 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \rightarrow (5) \quad A = \begin{pmatrix} k & 0 \\ k & 0 \end{pmatrix} \quad \rightarrow (6) \quad P_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

Synonymy refers to the case where there are more than one non-zero entries in a row of the matrix: thus, the same meaning is associated with two different words as shown in (7). However, in synonymy binomial sampling leads to an association matrix which is slightly asymmetric as shown in (8), which results in an asymmetric new P_1 matrix as shown in (9), which most likely gives rise to an even more asymmetric *A*, and synonymy decays over successive generations.

$$(7)\quad P_0 = \begin{pmatrix} 0{,}5 & 0{,}5 \\ . & . \end{pmatrix} \rightarrow (8) A = \begin{pmatrix} 1/2k+1 & 1/2k-1 \\ . & . \end{pmatrix} \rightarrow (9) P_1 = \begin{pmatrix} (1/2k+1)/k & (1/2k-1)/k \\ . & . \end{pmatrix}$$

The only stable solutions are given by (10) where synonymy has disappeared.

$$(10)\ P = (1 \quad 0) \qquad \text{or} \qquad (0 \quad 1)$$

We have seen stability of homophones in children's acquisition process. But the stability also depends on frequency of k samples. Our OE and ME data show that many of OE and ME homophones are high frequency words, i.e. words with large k samples, and low frequency words, i.e. words with small k samples became obsolete. In the real data of synonymy, *A* matrix and new *P* matrix are more asymmetric because they depend on frequency of words. Less frequent words became obsolete. Furthermore, the more near-synonyms in a given semantic field, the more became obsolete. However, new words enter and there are always competition and selection among near-synonyms (Ogura 1999).

6. Concluding remarks

We have investigated the evolution of homophony and its relation to the evolution of syllable number of words, based on the quantitative analysis on the historical data and simulation. Based on the CELEX database of English and the LDC Japanese Lexicon, we found that 22.8% of 52,447 types and 17.2% of 51,274 types are homophones in English and Japanese respectively. The simulation designed within the naming game framework shows that when the number of meanings and the number of utterances are equal, the agents converge to the same vocabulary, 20% of the words having homophones. We compared the quantitative studies with the simulation results and assumed that to avoid homophones, humans try to manifest "one meaning, one form", but homophones do occur and the threshold is around 20% of the vocabulary. We have suggested that homophones are the outcome of arranging form-meaning associations according to Zipf's law to maximize the referential power under effort for the speaker constraints. We have also discussed the neural bases of ambiguity and the relation between ambiguity and robustness.

Based on the CELEX database, the *Brooklyn – Geneva – Amsterdam – Helsinki parsed corpus of Old English*, and the *Penn – Helsinki parsed corpus of Middle English*, we found that homophones are stable and cumulate with the times. To avoid creating homophones, syllable number of words increases, with more recent entry dates of words associated with more syllables. Furthermore, we showed that larger mean syllable number of nouns than verbs is due to greater number of occurrence of nouns than verbs and their distribution of occurrence from 1 syllable to larger syllable number. Small syllable number of closed words is due to its small number of occurrence and the early date of entry. We also explored stability of homophones and instability of synonyms in children's association matrix based on the samples from parent.

The mechanism of the evolution of homophones and syllable length of words based on English data works cross-linguistically from the emergence of language and goes on at present. We have demonstrated that homophones and syllable length interact with each other to form complex adaptive system, and ambiguity, robustness and complexity are intertwined with each other in language evolution (Ogura – Wang 2004; Wang – Minett 2005).

* We wish to thank Michiyo Takeda for her assistance in data analysis. This research is supported by the grant from the Human Frontier Science Program.

REFERENCES

PRIMARY SOURCES

CELEX
1995 CELEX lexical database (version 2.5). Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen.

LDC
1997 LDC Japanese lexicon. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

The Brooklyn – Geneva – Amsterdam – Helsinki parsed corpus of Old English (2000).
The Penn – Helsinki parsed corpus of Middle English (2000).

SECONDARY SOURCES

Antilla, Raimo.
1989 *Historical and comparative linguistics*. (2nd revised edition.) Amsterdam – Philadelphia: John Benjamins.

Dossena, Marina – Roger Lass
2004 *Methods and data in English historical dialectology*. Bern: Peter Lang.

Ferrer i Cancho, Ramon – Ricard V. Sole
2003 "Least effort and the origins of scaling in human language", *Proceedings for National Academy of Sciences* 100: 788-791.

Greenbaum, Sydney – Geoffrey Leech – Jan Svartvik (eds.)
1980 *Studies in English linguistics*: *For Randolph Quirk*. London – New York: Longman.

Haiman, John (ed.)
1985 *Iconicity in syntax.* Amsterdam – Philadelphia: John Benjamins.

Harris, Alice C. – Lyle Campbell
1995 *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.

Huang, Churen – Winfried Lenders (eds.)
2004 *Computational linguistics and beyond.* Nankang: Academia Sinica.

Jespersen, Otto
1933 *Essentials of English grammar.* London: George Allen & Unwin.

Ke, Jinyun – James W. Minett – Ching-Pong Au – William S-Y. Wang
2002 "Self-organization and selection in the emergence of vocabulary", *Complexity* 7/4: 1-14.

Kempson, Ruth

1980 “Ambiguity and word meaning”, in: Sydney Greenbaum – Geoffrey Leech – Jan Svartvik (eds.), 7-16.

Koopman, Willem – Frederike van der Leek – Olga Fischer – Roger Eaton (eds.)

1987 *Explanation and linguistic change.* Amsterdam – Philadelphia: John Benjamins.

Lass, Roger

1980 *On explaining language change*. Cambridge: Cambridge University Press.

1987 “On sh*tting the door in Early Modern English: A reply to Professor Samuels”, in: Willem Koopman – Frederike van der Leek – Olga Fischer – Roger Eaton (eds.), 251-255.

Leech, Geoffrey

1974 *Semantics*. Harmondsworth, Middlesex: Penguin Books.

Maes Pattie – Maja J. Mataric – Stewart W. Wilson – Jean-Arcady Meyer (eds.)

1995 *From animals to animats 4. Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior.* Cambridge, MA: MIT Press.

Marchand, Hans

1969 *The categories and types of Present-Day English word-formation: A synchronic and diachronic approach*. (2nd edition.) Munchen: C. H. Beck’sche Verlagsbuchhandlung.

Nettle, Daniel

1999 *Linguistic diversity*. Oxford: Oxford University Press.

Nowak, Martin A. – Joshua B. Plotkin – David C. Krakauer

1999 “The evolutionary language game”, *Journal of Theoretical Biology* 200: 147-162.

Ogura, Mieko

1999 “Brain – language co-evolution in lexical change”. *Folia Linguistica Historica* XX/1-2: 3-23.

Ogura, Mieko – William S-Y. Wang

2004 “Dynamic dialectology and complex adaptive system”, in Marina Dossena – Roger Lass (eds.), 137-170.

Simpson, John – Edmund Weiner (eds.)

1994 Oxford English dictionary (OED). (2nd edition.) Oxford: Oxford University Press.

Quirk, Randolph – Sidney Greenbaum – Geoffrey Leech – Jan Svartvik

1972 *A grammar of contemporary English*. London: Longman.

Roberts, Jane – Christian Kay – Lynne Grundy

1995 *A thesaurus of Old English*. 2 vols. London: Centre for Late Antique and Medieval Studies.

Samuels, Michael L.

1987 “The status of the Functional Approach”, in: Willem Koopman – Frederike van der Leek – Olga Fischer – Roger Eaton (eds.), 239-250.

Steels, Luc

1996 “Emergent adaptive lexicons”, in: Pattie Maes – Maja J. Mataric – Stewart W. Wilson – Jean-Arcady Meyer (eds.), 562-567.

Strang, Barbara M. H.

1980 “The ecology of the English monosyllable” in: Sidney Greenbaum – Geoffrey Leech – Jan Svartvik (eds.), 277-293.

Wang, William S-Y. – Jinyun Ke – James W. Minett

2004 “Computational studies of language evolution”, in: Churen Huang – Winfried Lenders (eds.), 65-106.

Wang, William S-Y. – James W. Minett
2005 “The invasion of language: Emergence, change and death”, *Trends in Ecology and Evolution* 20/5: 263-269.

WordNet
2003 *WordNet* (version 2.0). Cognitive Science Laboratory, Prinston University.

Zipf, George Kingsley
1949 *Human behavior and the principle of least effort*: *An introduction to human ecology*. Cambridge, Mass.: Addison-Wesley.