Corpus Tools for Retrieving and Deriving Termhood Evidence

Chunyu Kit
Department of Chinese, Translation and Linguistics
City University of Hong Kong
E-mail: ctckit@cityu.edu.hk

1. Introduction

The necessity of integrating a set of corpus processing tools into a general architecture for computing in humanities has recently been recognised, as demand for corpus processing is growing beyond what existing tools can handle. It is pointed out that "…while there are now large collections of encoded texts, with the number continuing to grow rapidly, there has been little corresponding development of the 'next generation' of tools such as OCP, TACT and others that would enable humanities scholars to exploit fully the intellectual investment represented by these collections" [Sho99] and that "Many existing programs offer one or two transformation/ presentation tools (a KWIC display is an example), but few offer a rich enough set of tools to allow users to design their own display" [BH99].

What about corpus tools for terminological studies within the field of computational terminology? Many terminological works, like term bank construction, require proper judgment about the *termhood* of a term candidate, so as to ensure the quality of term identification. Such judgment requires adequate evidence, not surprisingly. However, what kinds of evidence do we need? Accordingly, what corpus tools shall we provide for retrieving, and deriving, such evidence from large-scale corpora in use for terminological studies?

Among many key issues in term identification – be it conducted by human experts or by automatic term recognition (ATR) programs, the most important is no doubt this one: How do we know a word sequence, of one word or more, is a term in a subject field? Since a term is known – actually, defined – to be a carrier of a key concept in a subject field, an equivalent question is: How do we know a word sequence carries a key concept?

Specialists of a subject field may not share the same criteria with terminologists (who may not be specialists) for term identification. The former are familiar with key concepts in their field, but may not have a good idea about how terms form and relate to each other, systematically, so as to represent concepts. The latter have good ideas about terms but may not have a thorough understanding of the key concepts in a subject field. Therefore, corpus tools are expected to provide necessary information to both terminologists and specialists about (1) how likely it is that a word sequence is a term, and (2) how important a concept – represented by a word sequence – is in a field. We may measure the importance of a concept in terms of its name's ranking in a thematic corpus representing a subject field. However, we would not over-expect any corpus tools – no matter how powerful they might be – to give a decent answer to a more subjective threshold problem like this one: How important does a concept have to be in its subject field in order to license its name to be a term?

This paper studies the kinds of evidence that are in need to support judgment about the termhood of term candidates, and propose necessary corpus processing tools for retrieving and deriving such evidence from large-scale corpora. The rest of the paper is organised as follows. The critical notions about termhood will be first introduced in Section 2, followed by a discussion in Section 3 on several kinds of termhood evidence. In Section 4, we will discuss corpus tools for retrieving and deriving termhood evidence based on suffix array, before concluding the paper in Section 5.

2. Termhood

Two critical notions are highlighted in [KU96] – a comprehensive review paper on methods of automatic terminology recognition. One is *unithood*, which "refers to the degree of strength or stability of syntagmatic combinations or collocations", and the other is *termhood*, which "refers to the degree that

a linguistic unit is related (or more straightforwardly, represents) domain-specific concepts".

The former is only concerned with term candidates of multiple words, and its measure (if there is one) indicates how likely it is that a term candidate is an *atomic* text unit. Actually, such *non-decomposability* of a multi-word term is strongly correlated to its representation of a concept. Since concepts are non-decomposable in human communication and terms represent key concepts in a subject field, any word sequence that does not function as an atomic linguistic unit in a language is known not to represent a concept and cannot, consequently, be a term. From this point of view, we can see the intrinsic relation of the unithood to the association of a linguistic unit to a domain-specific concept. In practice, the unithood can be understood as a criterion for filtering out non-terms – a way of selecting term candidates – for the next step of processing for term recognition, e.g., ranking with regard to their termhood.

Of course, ranking can be performed directly on the set of all n-grams of words from a corpus, i.e., the complete set of term candidates with no n-gram filtered out by any unithood measurement. This is another point of view from which we can see that the unithood is actually subsumed, in general, by the termhood. Therefore we can identify two senses for the term "termhood", one being its original sense as in [KU96] beyond unithood, and the other being the generalised sense as a comprehensive criterion to measure the likelihood of an n-gram being a term. It is in the generalised sense that we use it in the phrase "termhood evidence". However, this in no way means that the distinction of unithood and termhood is insignificant. To avoid confusion, simply bear in mind that when it is used in contrast to "unithood", it surely retains it original sense within that context.

In short, termhood indicates, in general, how likely it is that a word sequence is a true term. At first glance this might look like a paraphrase of the definition in [KU96]. However, a subtle difference lies in that KU's definition is assumed to apply only to term candidates after some unithood filtering, whereas the definition here applies to any possible candidate before any unithood filtering.

3. Kinds of evidence for termhood

Now, a question we want to ask is: What kinds of evidence do we want to dig out from a thematic corpus for the termhood of a possible term candidate? The answer depends, largely, on both the working procedure for term identification and on the measurements for the termhood. In general, there are two types of evidence, one *linguistic* and the other *statistical*.

The linguistic evidence for a term candidate includes (1) contexts of its occurrences, (2) its syntactic pattern, and (3) its semantic denotation. The last one is rarely available in the context of computational terminology; otherwise ATR would be a trivial task. To derive the syntactic pattern for a word sequence, a POS tagger would suffice – nowadays POS tagging technology has matured enough to meet this demand [Bri92, Ku92]. It is common practice to apply pre-defined syntactic constraints to filter out term candidates, e.g., only noun compounds are considered as multi-word candidates, and any word sequence containing a preposition is excluded. The strategies of utilising syntactic patterns in terminology processing are rather *ad hoc*, unfortunately, because the relation between any syntactic pattern and termhood is so uncertain. In contrast, the first type of evidence is more significant. It is to be retrieved by a concordance program, which is usually designed to retrieve a key word in context (KWIC), but can be easily extended to retrieve a sequence of key words in context. Since in many circumstances the contexts for a term candidate can be very large in number, how to utilise such contextual information with the aid of some statistical means remains a key problem in ATR. Simply displaying the contexts to a terminologist is of course better than none, but can we structure such information into a better format, or squeeze some significant statistical indication out from them? The number of contexts that a candidate can show up in provides the most important frequency information to compute the unithood and termhood, following the available measures that will be discussed below. Following the working procedure of [KU96], namely, filtering by unithood first and then ranking by termhood, we can group the statistical evidence (and relevant measures) into two categories for the convenience of discussion, namely, unithood evidence and termhood evidence.

3.1 Unithood evidence

A proper statistical measure for unithood quantifies the structural dependency between constituent words in a multi-word term candidate. An intuitive choice of measure for the dependency is the *co-occurrence frequencies* of two adjacent words, following the intuition that a higher co-occurrence frequency of two words indicates that they have a stronger dependency on each other. It can be formulated as follows:

$$d(w_i, w_j) \propto f(w_i, w_j) \tag{1}$$

where $d(\cdot,\cdot)$ denotes the dependency between two words and $f(\cdot,\cdot)$ their co-occurrence frequency. For two adjacent words, we have $j = i+1$.

However, there are counter examples to show that this measure is not always so reliable. For example, prepositions and determiners in English, e.g. "for the", "in a", co-occur very frequently, but they are not dependent on each other. An explanation for this is that there are many, if not more, occurrences of these words with other words. In contrast, the co-occurrence frequency of "Humpty" and "Dumpty", for example, is not high, but once one of them appears, the other always shows up as well. This shows that the proportion of one word's occurrences in the co-occurrences of the two is a better indication to the dependency. Accordingly,

$$d(w_i, w_j) \propto \frac{f(w_i, w_j)}{f(w_i)}, \frac{f(w_i, w_j)}{f(w_j)} \tag{2}$$

where $d(\cdot,\cdot)$ denotes the dependency between two words and $f(\cdot,\cdot)$ their co-occurrence frequency. Straightforwardly, each of the two items on the right-hand side is the estimation of conditional probability of a word given the other.

There are variants to combine the two items in the right-hand side of (2) into one. One choice is to take their product, or the log of their product, from which we can see its relevance to *mutual information* (MI), as follows:

$$d(w_i, w_j) \propto 2\log\frac{f(w_i, w_j)}{f(w_i)f(w_j)} \propto \log\frac{f(w_i, w_j)}{f(w_i)f(w_j)} \tag{3}$$

$$\mathrm{MI}(w_i, w_j) = \log\frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{4}$$

How close these two formulae are depends on the estimation of $p(\cdot,\cdot)$. It is conventional that $p(\cdot)$ is estimated by relative frequency through normalisation: $p(w) = f(w) / \sum_{w'} f(w')$. If we choose to normalise all frequencies with the same denominator, e.g., the total number N of all n-grams in a given training corpus, we may have $p(\cdot,...,\cdot) = f(\cdot,...,\cdot) / N$. Accordingly, we can rewrite (4) as follows:

$$\mathrm{MI}(w_i, w_j) = \log N + \log\frac{f(w_i, w_j)}{f(w_i)f(w_j)} \propto \log\frac{f(w_i, w_j)}{f(w_i)f(w_j)} \tag{5}$$

where log$N$ is a constant. An early application of MI in computational linguistics can be found in [CH90].

Probability is another option for measuring the dependency between constituent words in a word sequence. However, the factor of length needs to be taken into consideration, because longer sequences are in general less probable. It is unjustifiable to compare the absolute values of the probability of a long sequence and that of short one. *Perplexity* is a good choice for integrating these two factors, defined as in (6), for the comparison.

$$\mathrm{PP}(w_1...w_N) = P_{\mathrm{M}}(w_1...w_N)^{-\frac{1}{N}} \tag{6}$$

where $P_{\mathrm{M}}(\cdot)$ is the probability of a string estimated in terms of some language model M. It can be understood that $1/\mathrm{PP}(w_1...w_N)$ is the average probability, in a sense, over each word in the sequence. And log perplexity is the average number of bits to encode the sequence.

(7) $$\log \mathrm{PP}(w_1 ... w_N) = -\frac{1}{N} P_{\mathrm{M}}(w_1 ... w_N)$$

Another comprehensive, and complicated, measure is *log likelihood ratio* [Dun93], which considers not only the co-occurrence frequency of two words but also their non-co-occurring and "non-occurring" (in a sense) frequencies. Its formula is given as follows:

(8) $$-2\log\lambda = 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)]$$

where k_1 is the co-occurrence frequency of the two words in question, n_1 and k_2 are, respectively, the frequencies of the two words co-occurring with other words, n_2 is the frequency of other word pairs, $p_i = k_i / n_i \,(\text{for } i = 1,2)$ and $p = (k_1 + k_1)/(n_1 + n_2)$. $\log L(\cdot,\cdot,\cdot)$ is defined as in (9), following a binomial distribution:

(9) $$\log L(p, k, n) = k \log p + (n - k)\log(1 - p)$$

The log likelihood ratio can be applied to measure the strength of the association of two words. According to the sample output in [Dun93], however, more frequent words (including many function words) have higher log likelihood ratios. How to apply this statistic measure to quantify structural dependency of a word sequence remains an interesting issue to explore. Illustrations of using the log likelihood ratio for terminology identification can be found in [DGL94] and [Co95].

The application of a χ^2 statistic, namely, ϕ^2, to compute the *association* of a word pair in parallel texts is exemplified in [GC91] based on a contingency table of co-occurrence frequencies within aligned sentences:

(10) $$\phi^2 = \frac{(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)}$$

where the contingency table is given as

$$a = f(w_1, w_2), \qquad b = f(w_1) - f(w_1, w_2)$$
$$c = f(w_2) - f(w_1, w_2), \quad d = N - a - b - c$$

with N as the total number of aligned sentences. Clearly, this ϕ^2 score can be adapted as a measure for the strength of structural dependency between constituent words within a multi-word term candidate, e.g., let N be the occurrences of bi-grams involving none of the two words in a corpus and opt for a contingency table similar to the one in [Dun93].

The learning-via-compression approach to unsupervised lexical learning also provides an overall measure for the dependency within a token sequence, which is referred to as *description length gain* (DLG) [KW99, Kit00]. It calculates the number of bits that can be gained by extracting a candidate sequence out from a given corpus X as a lexical item (or unit).

(11) $$\mathrm{DLG}(x_{i...j} \in X) = \mathrm{DL}(X) - \mathrm{DL}(X[r \rightarrow x_{i...j}] \oplus x_{i...j})$$

where the rule represents a transformation throughout X by the extraction of $x_{i...j}$ and $\oplus$ indicates the concatenation of the transformed corpus and the extracted lexical candidate. $\mathrm{DL}(\cdot)$ can be calculated in a very simple manner following the *empirical entropy,* as follows:

(12) $$\mathrm{DL}(X) = -\sum_{x \in V(X)} f(x) \log_2 \frac{f(x)}{|X|}$$

where $V(X)$ is the alphabet (or vocabulary) in X, and $|\cdot|$ is the length of X as a string. However, the application of DLG to measure the unithood of a term candidate for terminology identification is yet an interesting topic for further exploration.

Interestingly, the *internal* structural dependency within a sequence may also have to do with the *external* factors, e.g., the number of distinct contexts in which it appears. Following Harris' idea for identification of morphemes half a century ago [Ha70], we can measure the independence of a sequence by the number of distinct contexts it appears in: the greater this number, the more independent the sequence is, and, correspondingly, the stronger the internal dependency among its constituents will be. In

a recent attempt [CDFZ02], this idea is re-invoked and formulated as the *accessor variety* (AV) criterion for Chinese word extraction, with the aid of some heuristics such as filtering functional Chinese characters. There is reason to believe that this criterion provides critical information about the unithood (and even termhood) evidence for terminology identification, although it is worth pointing out that the proportion of distinct contexts (to all contexts) in which a term candidate can show up also needs to be considered, in addition to the absolute value of AV.

The quantity of information that a term candidate carries (see formula (10) above) may also play an important role. If its distribution over individual documents is also considered, the *tf-idf* weighting scheme [SB88] that is popularly used in the community of information retrieval is also of particular significance. The weight for a term t in a document d (in a document set D) is defined as

$$(13) \qquad \mathrm{w}_{\mathrm{d}}(t) = -f_d(t)\log_2\frac{|D(t)|}{|D|} = f_d(t)\log_2\frac{|D|}{|D(t)|}$$

where $f_d(t)$ is t's frequency in d, known as *tf*, and the remaining part on the right-hand side is known as *idf*, with $D(t)$ denoting the set of documents each of which contains t. We can see the *idf*'s resemblance to the formula for empirical entropy [Sha48]. Since it uses *df*, instead of the term's frequency in the whole document set, the *idf* of terms actually gives the information that they carry for discriminating documents from each other. Thus, it can be understood as a quantification of a term's ability, in number of bits, to differentiate between documents. Although it is questionable whether the *tf-idf* scheme can be directly applied to termhood ranking for terminology identification, it surely provides a valuable lesson for us to learn from: the information that a term candidate carries is also an important indicator of its termhood.

3.2 Termhood evidence

The termhood evidence is expected to help us make the judgment about how likely it is that a term candidate will be a true term. A measure for the termhood indicates a ranking scheme for term candidates for the purpose of selecting the true terms from more to less likely ones. Various heuristics, strategies and criteria were used, as reviewed in [KU96] and [CBP99]. Recent ones include *C-value* and *NC-value* [FA96, FA97, MA99, FA99] and *imp* value [Nak99]. The application of *C/NC-value* involves various strategies of applying frequency information for term ranking, together with linguistic heuristics. The *imp* value is an attempt to make use of contextual information for the same purpose. In this sense, it is closely related to Harris' idea and the criterion of accessor variety, but in a different formulation.

However, it is also very critical for us to ask why we use one criterion instead of the others. Do we expect a criterion in use to reflect our understanding of "why terms are terms"? We know that terms represent prominent key concepts in a subject field. So, why simply find a proper measure for such *prominence* of a term candidate in a field (represented by a thematic corpus of an adequate size) for the purpose of term identification? Consequently, a follow-up question is: how do we know a term candidate is more prominent than others? Is it because it occurs more frequently? Or occurs with some typical linguistic pattern? Or carries more information? Or has a more stable internal structure? Or appears in more distinct contexts? Having a greater compression effect when it is extracted from a corpus? Do any of these criteria capture the essence of the prominence of a term?

We wish to rank term candidates with respect to their domain-specific prominence. In other words, why terms are terms in a subject field is only because of their outstanding status in that field, instead of in any other field, in particular not in a general balanced corpus. Following this line of thinking, it is highly feasible to take a comparative approach for term candidate ranking by comparing their ranks in a *thematic* corpus and a *background* corpus, namely, a general balanced corpus. Accordingly, the prominence of a term candidate can be measured by the difference of its (relative) ranks in the two corpora. What ranking scheme should be used for term candidates in a corpus is a key problem to be explored in our research. The simplest approach is ranking by frequency [Z35, Z49], and the next choice is by information [Sha48], for which there can be two alternatives: the average information carried by a term candidate, or the sum of information carried by all of its occurrences. Aside from this, we also

have other choices, including perplexity, *tf-idf* weighting scheme, DLG measure, etc., as discussed above. A number of statistical tests may be applied, including t_test, Z_test, χ^2 test, etc., for which more details can be found in a textbook like [O98].

There is reason to believe that this comparative ranking scheme, no matter what ranking scheme is used for word sequences in an individual corpus, should more reliably reflect the prominence of the term candidates selected by the above unithood filtering.

4. Corpus tools for retrieving and deriving termhood evidence

We have many alternative choices for the measurement of termhood evidence. What kinds of corpus tools do we need in order to retrieve such evidence from large-scale corpora?

First of all, we need a concordance program to extract all occurrences of a term candidate, together with contexts. There are many concordancers available, including OCP [HM79-80, HM87], TACT [Bra91, LBMSW96], WordCruncher, MonoConc, Wordsmith Tools, and Concordance – more information about these systems can be found in [Ho01] with useful URLs. However, simply a concordancer for extracting KWIC from a corpus is far from enough for our purpose, because the term evidence needs to be further derived from the (co-)occurrence and contextual information. Moreover, a concordancer would run too slowly on a large-scale corpus of tens of millions of words, if it extracts key words by scanning through a corpus word by word for every query. Some sort of indexing technique must be applied in order to maintain an acceptable running speed on very large corpora. *Suffix array* is one of the best of such techniques, which not only subsumes a concordance program by the nature of its indexing method but also supports the fastest n-gram counting known to date.

The suffix array as a data structure, functionally equivalent to a PAT-tree but more efficient in space complexity, was proposed in [MM90] for string searching and in [NM94] for n-gram counting for a large n and Japanese word extraction. [KW98] reports an elaborative implementation based on [NM94]. Recently, its application to deriving term frequency and document frequency was reported in [YC01]. The basic idea of suffix array can be exemplified in the following three stages, assuming that a corpus is a sequence of tokens (either words or characters) in an array:

1. *Indexing*: Let each index represent the suffix string from the index to the end of the corpus;
2. *Sorting*: Sort all indices in terms of the alphabetical order of their suffix strings, resulting in all identical n-grams of any length lined up adjacent to each other;
3. *Counting* (or *retrieving* and *displaying*): Count the n-grams of any length by going through the sorted indices, or retrieve and display the concordance of a target n-gram.

The result of the sorting yields the concordance for any n-gram. The only thing that needs to be done in order to extend it into a concordancer is a binary search to locate a target n-gram in the sorted suffix array and display its occurrences with contexts. A great advantage of suffix array over PAT-tree is the simplicity in retrieving the left and the right context of each occurrence of any target n-gram. There is no need to construct a "prefix array" for retrieving left contexts.

When the suffix array of a given corpus has been sorted and all n-grams counted, there are two more tasks to carry out towards the purpose of deriving termhood evidence to facilitate terminology identification: (1) computing the termhood score(s) following some of the above measures picked by a user, and (2) displaying the computing results in a proper way of visualisation. The visualisation of the difference of a term candidate's ranking in a thematic corpus and a background corpus is no doubt one of the most important tasks for (2).

5. Conclusions

In this paper, we have discussed the critical notions of termhood, including the original notions of unithood and termhood, and presented various kinds of termhood evidence to facilitate terminology identification, including linguistic and statistical evidence. The unithood is argued to be a criterion for term candidate filtering, and the termhood for candidate ranking. Many statistical measures for scoring unithood have been discussed, including mutual information, perplexity, log likelihood ratio, ϕ^2 score,

description length gain, and the accessor variety criterion. The idea of ranking term candidates with respect to their prominence in a thematic corpus (representing a subject field), in contrast to that in a general balanced background corpus, has been proposed. Accordingly, a number of ranking criteria for this purpose were discussed, including occurrence frequency, the amount of information that a candidate carries, and the *tf-idf* weighting scheme borrowed from IR. The effectiveness of these criteria needs to be examined through experiments with real data. Finally, suffix array and its easy implementation are presented as a basis not only for n-gram counting but also for retrieving and deriving termhood evidence for the purpose of facilitating terminology identification work.

Acknowledgements

The author wishes to thank Randy LaPolla and Robert Neather for their helps to improve this paper. The author is of course responsible for remaining imperfections.

References

[BH99] J. Bradley and T. Horton. 1999. Text Analysis Tools: Architectures and protocols. *ACH-ALLC'99*, June 9-13, University of Virginia.

[Bra91] J. Bradley. 1991. *TACT User's Manual*. University of Toranto.

[Bri92] E. Brill. 1992. A simple rule-based part-of-speech tagger. In *Proc. of 3rd ANLP*, pp.152-155, Trento, Italy.

[CDFZ02] K. Chen, X. Deng, H. Feng, and W. Zheng. 2002. Accessor variety criterion for Chinese word extraction. Manuscript, Dept. of CS, CityU of HK.

[CH90] K. W. Church and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1):22-29.

[CHH91] K.W. Church, P. Hanks and D. Hindle. 1991. Using statistics in lexical analysis. In U. Zernik (ed.), *Lexical Acquisition: Exploiting one-line resources to build a lexicon*, Erlbaum.

[Chi97] L. Chien. 1997. PAT-tree-based keyword extraction for Chinese information retrieval. *ACM SIGIR'97*, pp.50-58.

[Co95] J. D. Cohen. 1995. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, **46**(3):162-174.

[DGL94] B. Daille, E. Gaussier amd J.-M. Langé. Towards automatic extraction of monolingual and bilingual terminology. *COLING'94*, pp.515-521.

[Dun93] T. E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1):61-74.

[FA99] K. T. Frantzi and S. Ananiadou. The C-value/NC-value domain independent method for multiword term extraction. *Journal of Natural Language Processing*, **6**(3):145-180.

[GC91] W.A. Gale and K.W. Church. Identifying Word Correspondences in Parallel Texts. In *Proceedings of the 4th Speech and Natural Language Workshop*, pp. 152--157. DARPA, Morgan Kaufmann.

[Ha70] Z. S. Harris. 1970. *Papers in Structural and Transformational Linguistics*. Dordrecht, Holland: Reidel

[HM79-80] S. Hockey and I. Marriott. 1979-80. The Oxford concordance project (OCP), parts 1-4. *Literary and Linguistic Computing*, 7:35-43, 155-164, 268-275; 8:28-35.

[HM87] S. Hockey and J. Martin. 1987. The Oxford concordance program version 2. *Literary and Linguistic Computing*, 2:125-131.

[Ho97] S. Hockey. 2001. Concordance programs for corpus linguistics. In R. C. Simpson and J. M. Swales (eds.), *Corpus Linguistics in North America*, pp.76-97. University of Michigan Press, Ann Arbor.

[Kit94] C.Y. Kit. 1994. Automatic terminology extraction for thematic corpus based on sub-term co-occurrence. In *ROCLING-V*, July, Taiwan.

[Kit00] C. Y. Kit. 2000. *Unsupervised Lexical Learning as Inductive Inference*. PhD thesis, University of Sheffield.

[KU96] K. Kageura and B. Umino. Methods for automatic term recognition: A review. *Terminology*, **3**(2):259-289.

[Ku92] J. Kupiec. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, **6**:225-242.

[KW98] C. Y. Kit and Y. Wilks. 1998. The Virtual Corpus approach to deriving n-gram statistics from large-scale corpora. In C.N. Huang (ed.), *Proceedings of 1998 International Conference on Chinese Information Processing*, pp.223-229, Beijing.

[KW99] C.Y. Kit and Y. Wilks. 1999. Unsupervised learning of word boundary with description length gain. In M. Osborne & E. T. K. Sang (eds.), *CoNLL-99*, pp.1-6. Bergen, Norway, 12 June.

[LBMSW96] I. Lancashire. J. Bradley, W. McCarty, M. Stairs and T. R. Wooldridge. *Using TACT with Electronic Texts*. Modern Language Association of America, New York.

[MM90] U. Manber and E. Myers. 1990. Suffix array: a new method for on-line string searches. In *First ASM-SIAM Symposium on Discrete Algorithms*, pp.319-327, American Mathematical Society, Providence.

[NM94] M. Nagao and S. Mori. 1994. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. *COLING-94*, pp.611-615.

[Nak99] H. Nakagawa. 1999. Experimental evaluation of ranking and selection methods in term extraction. In D. Bourigault, C. Jacquemin and M.-C. L'homme (eds.), *Recent Advances in Computational Terminology*, pp.303-325, John Benjamins.

[O98] M. P. Oakes. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

[SB88] G. Salton and C. Buckley1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**:513-523.

[Sha48] C. Shannon. 1948. A mathematical theory of communication. *Bell System Tech. J.*, **27**:379-423, 623-656.

[Sho99] H. Short. 1999. Text analysis tools: Architectures and protocols. *ACH-ALLC'99*, June 9-13, U. of Virginia.

[SST98] M. S. Sun, D. Y. Shen and B. K. T'sou. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. *COLING-ACL'98*, pp.1265- 1271.

[YC01] M. Yamamoto and Church, K. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, **27**(1):1-30.

[Z35] G. K. Zipf. 1935. *The Psycho-Biology of Language*. Houghton Mifflin.

[Z49] G. K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. Hafner.