2005

The head-modifier principle and multilingual term extraction

Andrew R. Hippisley
University of Kentucky, andrew.hippisley@uky.edu

David Cheng

Khurshid Ahmad

Repository Citation

Digital Object Identifier (DOI)
10.1017/S1351324904003535

The head-modifier principle and multilingual term extraction[1]

Andrew Hippisley a.hippisley@surrey.ac.uk

David Cheng d.cheng@surrey.ac.uk

Khurshid Ahmad k.ahmad@surrey.ac.uk

Department of Computing
School of Electronics and Physical Sciences
University of Surrey
Guildford, Surrey, UK
GU2 7XH

Abstract

Advances in Language Engineering may be dependent on theoretical principles originating from linguistics since both share a common object of enquiry, natural language structures. We outline an approach to term extraction that rests on theoretical claims about the structure of words. We use the structural properties of compound words to specifically elicit the sets of terms defined by type hierarchies such as hyponymy and meronymy. The theoretical claims revolve around the *head-modifier* principle which determines the formation of a major class of compounds. Significantly it has been suggested that the principle operates in languages other than English. To demonstrate the extendibility of our approach beyond English, we present a case study of term extraction in Chinese, a language whose written form is the vehicle of communication for over 1.3 billion language users, and therefore has great significance for the development of language engineering technologies.

1 Introduction

Natural language processing (nlp) and natural language engineering (nle) systems operate on natural language texts whose structures e.g. discourse structure, clauses, phrases, and words are the objects of theoretical linguistics. The connection between nlp/e and linguistics has seen clear benefits for linguists where systems have been designed to allow them to evaluate their theories. These systems demonstrate ‘the instrumental use of computation in the pursuit of linguistic goals’ (Thompson 1983: 23), early examples of which are the parsers developed for Generalized Phrase Structure Grammar, and a more recent example of which is the DATR lexical knowledge representation language (Gazdar and Evans 1996) used to validate Network Morphology theories (Corbett and Fraser 1993, Hippisley 2001). It has also been argued that nlp/e can benefit from insights based on theoretical studies of language. In information retrieval / extraction there are attempts to enhance simple string-based methods by considering the grammatical structures in which key words appear in order to "uncover certain critical *semantic* aspects of document content" (Strzalkowski et al. 1999: 113). One of these structures is the compound noun which has received attention first because the overwhelming majority of key word are nouns, and second because most of these are multi-word terms. Linguistic insights into the semantic interpretation of these structures could be used to "uncover" document content conveyed by multi-word terms. The particular insight we consider is the head-modifier principle.

Sparck Jones (1985) in an early paper on compound nouns in nlp pointed to three interpretation challenges associated with noun compounds: bracketing, the exact meaning of the compound's constituents, and the interpretation of the relationship between the constituents. She observes that any solutions to the first two would have to be based on general tendencies. And only the third, the relationship between the elements of a compound, can be grounded on a principle which is claimed to be universal, namely the head-modifier principle. In a compound word consisting of two or more elements, it is claimed that the linear arrangement of the elements reflects the kind of information being conveyed. One element, identified as the head, acts to name the general (semantic) category to which the whole word belongs; other elements, modifiers, distinguish this member from other members of the same category. In this way the head-modifier principle identifies a set of terms related through *hyponymy* with the head of the compound constituting the hypernym. In a construction such as *houseboat* the head element is *boat*, and can therefore be viewed as the hypernym. The compound *houseboat* it therefore a hyponym of boat, i.e. a kind of boat. The modifier *house* acts to distinguish this member from the other members of the set of hyponyms, for example another hyponym is *speedboat*. This is shown as a *type hierarchy* in Figure 1.

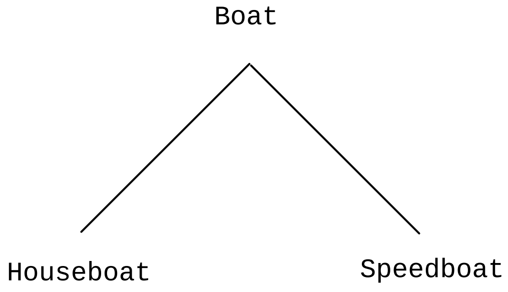

Figure 1. Compounds and hyponymy

Important for information retrieval / extraction is the fact that this is a domain independent principle which can be used to extract content from domain dependent objects. Further there is a second sense in which it is domain independent: because of its claimed universality in the structure of words in natural languages it can be employed in systems operating over texts other than English. The head-modifier principle has been used in language engineering tasks for a number of languages. We show its use in Chinese term extraction as an example of its use beyond English texts.

Section 2 is a brief discussion of the head-modifier principle and its role in compounding, We illustrate with data from both English and Chinese to underline its universality as a linguistic principle, and its applicability to information retrieval / extraction in more than one language. In section 3 we give an overview of how the head-modifier principle has been employed in a variety of information tasks, including automatic term recognition, lexicon induction, query refinement and term conflation. Section 4 outlines in detail the application of the head-modifier principle outside English, i.e. to Chinese term extraction, showing how it can be used to extract term sets associated by the thesaural relations of hyponymy and meronymy in the domain of information technology, part of the Chinese lexicon experiencing particularly rapid growth.

2 The head-modifier principle and multi-word terms

Terminologists such as Felber have observed that major developments in all fields of human endeavour during the 20th century have led to an influx of millions of concepts, but that there is a deficit of terms to name them: ‘All these concepts have to be represented by terms in individual languages which have a restricted word and word element stock for term formation.’ (Felber 1984: I). There are three main ways open to a language to expand its term stock. One is simply to borrow from a source language which has already associated the given concept with a term. The introduction of the term into the language’s lexical stock can be insensitive to differences in grammatical structure between the source and target languages, including morphotactics and phonotactics, and this can lead to the term’s ultimate rejection. The second way is to find translation equivalents of the source term so that the borrowed term is structurally native to the language. With multi-word terms, which is the majority, equivalents must be determined for all the constituents. These are *loan translations* in Haugen’s (1950) taxonomy of borrowed terms, and they are a major means of term stock expansion. The third way is entirely language-internal: a new term is created from the resources of the target language to designate the new concept. A number of authors, including Rogers (1997) and Heid

(1999), have remarked on the productive use made by special languages of word formation operations available in the target language to derive new terms from existing lexical items. Two important word formation operations are *affixation* and *compounding*.

One of the ways in which languages differ is their preference for a specific kind of word formation operation. A *fusional* language like English uses both affixation and compounding. On the other hand *isolating* languages such as Chinese have few affixes and make almost exclusive use of compounding (see for example Anderson 1985). This is illustrated in Table 1.

Table 1. *Word formation operations in two typologically distinct languages*

Fusional Language: English	Operation	Isolating Language: Chinese	Operation
process → processor	affixation	處 理器 *chŭ-lĭ qì* process tool 'processor'	compounding
processor → mobile processor	compounding	流 動 處 理 器 *líu-dòng chŭ-lĭ qì* mobile process tool 'mobile processor'	compounding

From the table we can note that where English uses affixation to derive a word, e.g. *processor*, compounding is used for the Chinese equivalent. But compounding is a word formation operation productively used by both typologically distinct languages, e.g. English *mobile processor* and its Chinese equivalent *líu-dòng chŭ-lĭ qì,* literally 'mobile processing tool'. A good working definition of compounding is provided by Trask, and is consistent with all our compound examples:

> ‘The process of forming a word by combining two or more existing words: *newspaper*, *paper-thin*, *babysit*, *video game*.’ Trask (1993: 53).

From Trask’s examples it should be noted that orthographically a compound in English is represented with or without a space between constituents, and sometimes with a hyphen. A test to determine whether two words are actually elements in a compound comes from the fact that compounds have one primary stress, a property of all words[2]. In Chinese, which is our main focus, the writing system does not distinguish word boundaries hence there is no spacing between morphemes, including constituents of a compound. Moreover from Trask’s definition compounds may be combinations of more than two existing words; compounds consisting of three, four and five elements will feature in our discussion.

Since English and Chinese both make extensive use of compounding for creating new terms, for both languages the universal head-modifier principle must play an important role in term formation.

2.1. The head-modifier principle in compounding

The notion of head and modifier is inherent to many grammatical descriptions. It is assumed in Dependency Grammar, X-bar grammar, Generalized Phrase Structure Grammar, Head-Driven Phrase Structure Grammar and in Word Grammar approaches to the lexicon (see for example Bauer 1994, Fraser, Corbett and McGlashan 1993 and Zwicky 1985 for details). In a syntactic construction one of the constituents acts as the head, or core of the phrase, and the other constituents as dependents on it, or modifiers of it. There is a default association between the syntactic head, and the core semantics of the phrase. In nlp, automatic parsers make use of this default association. For example Abney's (1991) parser converts a stream of words into semantically based phrase-like units called chunks. The content word falling in syntactic head position within the chunk specifies the semantic head in which the chunk is rooted.

Heads are also a powerful descriptive device in the lexicon, namely in compound formation[1]. Consider the following examples (based on Spencer 1991: 310).

(1) [film society]

(2) [[film society] committee]

(3) [[[film society] committee] scandal]

In (1) *society* is modified by *film*: the rightmost element is the head of the construction, the element to the left is the modifier of the head. The head-modifier relationship is important for semantic interpretation in that 'the meaning of the construct is a sub-type of the head' Zwicky (1993: 296). Thus film *society* is a type of a *society*. At the same time, the modifier plays a 'contributory role, restricting the meaning of the head in one way or another.' Of all the possible *societies* the head could be denoting, the modifier acts to pin it down to denoting the 'film' type. In this way heads and modifiers

[1] Heads also play a role in affixal word formation. Which constituent, the affix or the stem, is viewed as the head has been a matter of debate. For heads as affixes, see Williams (1981); for heads as stems, see Beard (1998: 50-53).

express hyponymy relations between lexical items. In (2) we see the original compound in (1) acting now as the modifier of a compound whose head is *committee*; the compound in (2) then functions as modifier of a new compound in (3) where the head is *scandal*. From the examples we should note (at least) two formal properties associated with heads in English compounds. First, their position is consistent: they always appear at the right edge of the construction. Secondly, the properties of the head determine the syntactic category of the entire construction. Note that the bracketing is important in the examples as it indicates the subconstituency of the compound, and therefore its derivational history. The internal brackets express the origin or *root* of the compound. In (2) we have a compound where material has clearly been added to the right of the expression *film society*, i.e. the added material is located at the head of the new compound. In (3) the added material is also located at the head of the new compound whose origins are the expression *film society committee*.

The bracketing in (2) and (3), in combination with the head-modifier principle, indicate how to interpret the compound. But without the bracketing, which of course is the standard situation, there is more than one intepretation. The example in (2) has the alternate bracketing, shown in (4), where the head is part of a pre-existing two element compound modified by *society*.

(4) [film [society committee]]

The interpretation of (4) is something like: “There exist committees, some of which are society committees. There are range of these, including society committees whose interest is film.” Recall from section 1 that resolving bracketing ambiguities was amongst

Spark Jones' list of the nlp compounding challenges. Text frequencies have been used to help resolve these ambiguities. It has been observed that right-branching compounds, as in (4), are generally much rarer (see Lauer 1995), and this is usually factored into diambiguation algorithms. Text frequencies of the bracketed elements are also used. For example, the frequency of [society committee] in (4), which should be zero occurrences, can be compared with that of [film society] in (2) to give the likelihood of the candidate bracketings. Lieber and Sproat (1992) note that stress plays an important disambiguating role. Using the Compound Stress Rule from Chomsky and Halle (1968) the claim is that in left-branchng compounds (examples (2) and (3)) stress is on the first element; the example they give is *Áir force* academy, which has the bracketing [[Air force] academy]. In right-branching compounds stress in on the middle element: *radio diréction finder* has the bracketing [radio [direction finder], i.e. a type of direction finder. These prosodic properties are important for speech processing applications (Sparck Jones 1985: 376).

2.2 Using the head-modifier principle to query multi-word terms

The head-modifier principle that is claimed to underlie compound formation can be used as the basis of two simple pattern matching schemas to elicit terms and the thesaural relations between them. One is constructed to elicit the members of a given category based on the hyponymy relation between words, and the other possible attributes associated with a category member based on the meronymy relation between words. We begin with the first schema, given in (5). Its target strings are the set of compounds whose head element is equivalent to the substring in the query, hence in bold. Following the head modifier principle, what distinguishes one target string from another is located in the modifier element. From the bracketing we see that the query substring marks the root of a compound where new material is found to the left.

(5) [X_N [**substring**]]

Given the term *boat* we may extract the set of its hyponyms by using a query that fits the schema in (5). Table 2 shows how we search for the set of strings representing the hyponyms of *boat*.

Table 2. *Eliciting hyponyms of* boat

Information extraction task:	Search schema used:	Query example:
Elicit hyponyms of term named in query	[X $_N$[**substring**]]	[X $_N$[**boat**]]

The possible results of the query are shown in Table 3 where the target strings constitute the set of hyponyms of a term named in the query string, in this case *boat*. For each target string the modifier element is a noun. Note how this element acts to distinguish one hyponym from another.

Table 3. *Elicited hyponyms of query string* boat

query string	Sample target strings
[X [**boat**]]	house**boat** speed**boat** river**boat**

An important aspect of the search schema is that it can be used recursively. A target string in Table 3 can itself be the head of a compound. It therefore supplies the query substring of a new query. Items recovered from this new query will represent hyponyms of a term which is itself a hyponym of a previous query. The target string *speedboat* of Table 2 can occupy the head position in a new query: [X_N [**speedboat**]]. Again what is being queried is the set of strings which consist of a noun plus the string *speedboat*. Sample target strings could include *competition speedboat* and *leisure speedboat*. It should be noted that in order to retrieve a compound term based on an already existing term we must make reference to part of speech tags. In Table 2 target strings are collocations of any string that belongs to the class of nouns followed by the string *speedboat*. The assumption is that collocations of noun plus noun constitute noun-noun compounds where the rightmost noun is the head element (see discussion in §3 for other work making this assumption). Headed compounds in English are typically noun-noun compounds. The hyponymy relations between the extracted terms is graphically represented in Figure 2 where members of a category can act as sub-categories which themselves have members. Each level of the hierarchy is related to a query. Different query strings are used to elicit the second and third levels of the hierarchy, as shown by the different numerical subscripts.

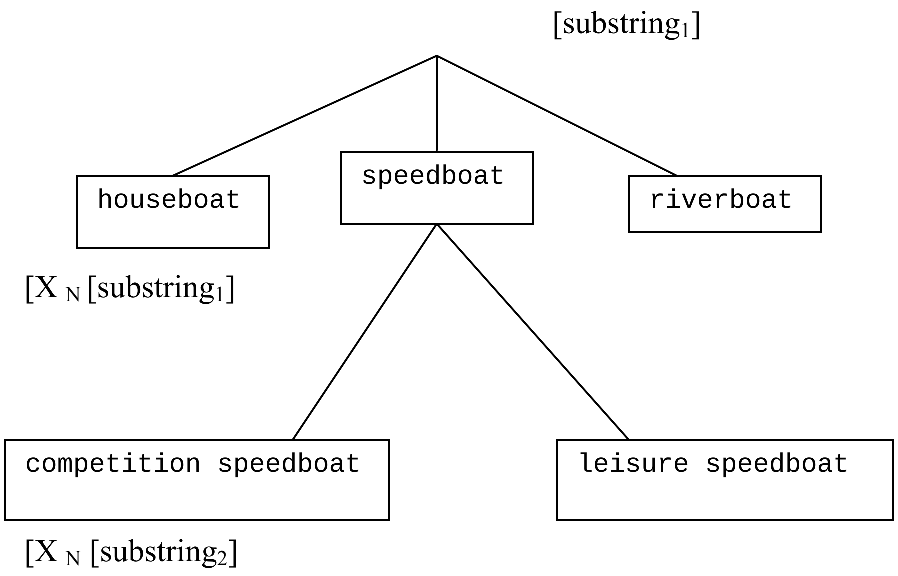

Figure 2. Type hierarchy elicited by the first head-modifier search schema

It will be noted that this first search schema is used to elicit words associated through the hyponymy relation: *speedboat* is a hyponym of *boat*, and *competition speedboat* is a hyponym of *speedboat*, etc. The second schema allows for a different task: it is used to elicit possible attributes of a term. In this way it elicits words associated by the *meronymy* (part-whole) relationship.

(6) [[**substring**] X $_N$]

The second schema can be used to redirect the focus of a query from eliciting hyponyms to eliciting attributes of a given term. This is achieved by using target strings of the first schema to form the query substring of the second schema where what is being queried is the set of compounds that share a common modifier. The serial application of the two schemas is important: where the first schema extracts compounds that represent hyponyms of a given (hypernym) term, the second schema extracts terms representing attributes of the compound, i.e. its *meronyms*. This is shown in Table 4.

Table 4. Eliciting meronyms of the extracted term speedboat

Information extraction task:	Search schema used:	Query example:
Elicit meronyms of extracted term	[[**substring**] X_N]	[[**speedboat**] X_N]

For this query, the possible results will be attributes of the term found in the query string, i.e. the meronyms of a term named in the query. This time target strings represent a set of compounds distinguished not by the modifier element but by the head element. This element acts to name the attribute of a term expressed by the modifier element.

Table 5. Elicited attributes of Speedboat

Query string	Sample target strings
[[**speedboat**] X_N]	**speedboat** length **speedboat** size **speedboat** engine

We can summarise the tasks of both search schemas as follows. Given the initial query string *boat* the first schema elicits terms that are hyponyms of *boat*, including *speedboat*. The second search schema elicits attributes or properties of the elicited term, i.e. its meronyms, such as *speedboat length*. The set of terms extracted is represented hierarchically in Figure 3. It should be noted that the hierarchy represents a hyponym relationship between the root node and its daughter node, expressed by a solid line, but a meronymy relationship between the daughter node and its daughters, expressed by a broken line. The queries used to elicit each level of the hierarchy are clearly shown.

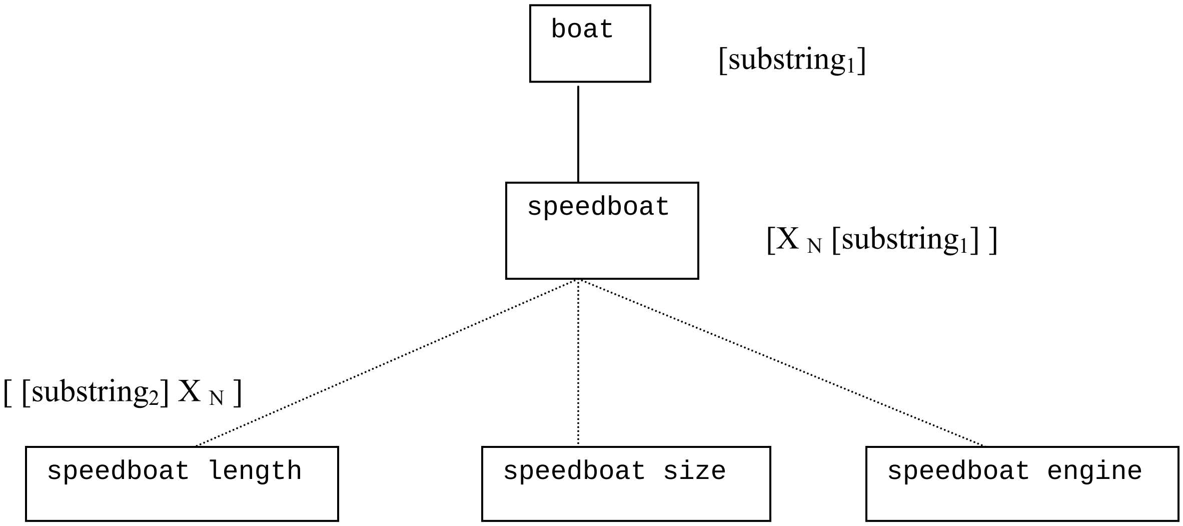

Figure 3.Type hierarchy elicited by the serial application of the first and second schemas

3. Applying the head-modifier principle to natural language engineering

The head-modifier principle has found its way into a number of information retrieval / extraction techniques as one of a number of means of accessing document content. It has been used as a 'bridge' between explicit, detectable syntactic constructions and the implicit semantics embedded within them. Ruge (1997) makes this point well:

> "Head modifier relations bridge the gap between syntax and semantics. On the one hand they can be extracted on the basis of pure syntactic

> analysis. On the other hand the modifier specifies the head, and this is a semantic relation."

In this section we briefly review the role it has played in three related areas: the (semi)automatic induction of semantic lexicons, the identification of technical terms in a corpus, and query refinement. In each approach there is a shared aim to find structures in what is perceived to be the largely unstructured text resource of text corpora. Machine-readable dictionaries (MRDs) represent a much more structured text resource but have been found to be unsatisfactory in completeness, and in consistency in the way the lexical knowledge is represented (e.g. Hearst 1992; Boguraev and Pustejovsky 1996: ch. 1). The move away from the pre-encoded knowledge offered by MRDs towards a "knowledge poor" resource such as free text (Grefenstette 1994: 17) requires the cataloguing of repeated structures, such as those defined by the head-modifier relation, with the aim of uncovering the knowledge embedded in them. A good example of this is Hearst (1992) which identifies half a dozen repeated 'lexico-syntactic patterns' in free text that embed the hypernym-hyponym relation between terms. One of these is covered by the regular expression in (7).

(7) NP {, NP}* {,} or other NP

The claim is that the NP on the right hand side of 'other' will be the hypernym of which NPs to the left-hand side are hyponyms, as in the example "Bruises, wounds, broken bones (hyponyms) or other injuries (hypernym)". We begin with how the head-modifier can assist in generating from text specialist semantic lexicons used in many nlp tasks. .

3.1 Inducing semantic lexicons

Boguraev and Pustejovsky (1996) stress the importance of the computational lexicon in nlp systems, and point to electronic text corpora as a possible source from which a lexicon can be (semi) automatically derived. A corpus of specialist texts will yield a semantic domain-specific lexicon. Riloff and Shepherd (1999) describe an algorithm that automatically induces semantic lexicons of specialist fields by exploiting constructions that specify some sort of semantic relation between the construction's components. A limited number of hand-picked core terms, or seed words, act as representatives of a semantic class. These are then retrieved from a chosen set of grammatical constructions, along with the other words appearing in the construction. These other words are added to the semantic class of the seed word based on the assumed semantic affinity of elements of the same construction. This process is iterative as newly added words become the seed words for the next search. One of the four constructions identified is noun-noun compounds where the head-modifier principle is used to suggest hypernym-hyponym relations between the seed word and the other elements in the construction. An example from the results data is the seed word *bomb*, representative of the Weapons semantic class, which picks out *car_bomb*, a type of bomb, and correctly adds it to the Weapons class. An approach to lexicon acquisition from free text that uses the context of repeated constructions not only to assign semantic class but also to specify the full set of semantic feature values of a lexical item is that of Pustejovsky et al. (1993). In this approach, based on a full blown theory of lexical semantics, the seed words come with a partially specified lexical semantic structure, the set of *qualia*, that is inferred from MRD representations. Without going into the details of the approach, one of the processes involves the induction of taxonomic relations through headed noun-noun compounds. In this way the qualia of the head noun will be shared, and further specified, by the modifier.

A final illustration of the use of the head-modifier principle in the induction of lexicons from corpora comes from Soderland et al (1995). They describe a system for generating conceptual dictionaries from specialist texts for use by information extraction systems. The dictionary consists of a number of abstract case frame definitions, each being a set of filled and unfilled semantic and syntactic slots. The unfilled slots for a case frame definition are filled by noun phrases satisfying phrasal constraints specified in the definition. For some of the definitions the constraint on suitable material is partially based on the head-modifier principle, as shown in for Prepositional Phrase constraint in (8).

(8) CN-type: Diagnosis
Subtype: Pre-existing
Extract from Prep.Phrase "WITH"
Passive voice verb
Verb constraints
    words include "DIAGNOSED"
Prep. Phrase constraints:
    preposition = "WITH"
    words include "RECURRENCE OF"
    modifier class <Body Part or Organ>
    head class <Disease or Syndrome>

This case frame is used for the class of Diagnosis sentences, and the sub-class of Pre-existing diagnoses, and will extract sentences such as "…diagnosed with recurrence of *lung cancer*", the italics indicating the extracted information which is a headed noun-noun compound. The unfilled slots require material tagged as 'Body Part or Organ' and

‘Disease or Syndrome’. But more than that it must be arranged with the ‘Disease or Syndrome’ appearing on the right hand side, as the head of the compound, according to the head-modifier principle. The use of the principle is made explicit by the labeling of the slots ‘modifier class’ and ‘head class’.

3.2 Identifying technical terms

As mentioned in section 1, many technical terms are multi-word, i.e are compounds or phrases. For example Justeson and Katz (1995) claim that the majority of technical terms are nominal compounds, based on searching through a range of technical dictionaries. This is because single words are usually polysemous and modification of an existing noun through compounding narrows down its possible interpretations, a fundamental requirement in terminology (Sager et al. 1980: 268). It is therefore not surprising to find the head-modifier principle playing a role in term identification systems. Justeson and Katz (1995) propose a term identification algorithm which makes partial reference to the head-modifier principle. They observe that compound terms have different properties to ordinary compound words. One of these properties concerns the tendency to omit the modifier in subsequent uses of the compound. They argue that the tendency is much stronger in ordinary compounds since word sense can be inferred from the head noun alone, and much weaker in specialised compound terms where the specificity of a term requires the presence of all its surface elements. This property can be used in assisting to distinguish terms from ordinary words. Frantzi and Ananiadou (1997) in their automatic term recognition algorithm assume multi-word noun terms to be the default, following

Sager et al. (1980). Their linguistic filter for extracting terms is the basic constituent structure of a right-headed compound noun.

The related area of phrase normalization has as its starting point the fact that mult-word terms are more useful as representatives of semantic content in a text than single word terms (e.g. Strzalkowski et al. 1999). A term will usually consist of more than one content word, as in a compound. But also there will exist in a text paraphrases of the term in the form of various syntactic constructions. Sager et al. (1980) for example show that process compound terms, such as *temperature control*, have parallel syntactic constructions, i.e. *control of temperature*. If the set of variants can be traced in the text then multi-word terms can be conflated for indexing, in much the same way as single word terms are conflated through stemming. This is possible since the paraphrases of a compound term are limited to a narrow range of constructions involving the elements of the compound. Identifying variants is a matter of searching syntactic patterns that contain the content words of the multi-word term. Thus the paraphrase of *control of temperature* has the pattern [NP_1] [P] [NP_2] where [NP_1] = head element of the compound; [P] = *of*; and [NP_2] = modifier element. In examples such as these term conflation becomes a matter of finding and matching head and modifier pairs in the text for each term, e.g. matching the head *control* for *temperature control* and *control of temperature*. This process is integral to the natural language information retrieval system described in Strzalkowski et al. (1999). A tagged text parser generates simple parse trees of clauses. These express head-modifier relations, and parses which have the same head-modifier relations are conflated. A similar approach is used in Evans et al. (1991). They then compare the elicited list of terms with a ‘certified terminology’. An exact match confirms a string as a term. But they use compound structure to suggest that an elicited term that is a substring of a certified term is a more general instance of it, and a string including a certified terms is a more specific instance of it.

3.3 Query refinement

In IR it is well known that queries based on a single word result in poor recall and precision rates. This is because most words are highly polysemous, so that the user may have one meaning in mind but documents with all possible meanings will be retrieved. Grefenstette (1997), amongst others, notes that the ideal is long descriptive queries, yet that this falls short of the reality: the 'typical' user inputs extremely short length queries, unaware of the single word polysemy. One way of bringing the reality closer to the ideal is to refine a user's initial query by automatically locating and presenting the full range of meanings of the single word query, with required disambiguating textual information. The user can select the word and its context and re-run the query more successfully. This is viewed as an intermediate structure, sitting between single word query and the texts, and Grefenstette outlines a number of techniques for automatically generating such structures. The main idea is to pinpoint structures in which the word appears, and infer the particular meaning of the word from the structure in which it is found. The structure of nominal compounds can be used in this way. Sager et al. (1980) note that frequent words tend to have low information value, presumably due to high polysemy, and are therefore the items that are most frequently modified. If it is the heads of compound nouns that are ambiguous, then the modifier will provide the appropriate disambiguating context. Greenstette provides regular expressions to act as headed noun filters. For example (9) picks out the compound *red warning lights* :

(9) (PRE)* NOUN

where the PRE class specifies modifiers, and is defined by nouns (amonst other parts of speech) and the NOUN class is the head, defined as singular and plural nouns. One example given is the single query *watch*. Where *watch* is found in the NOUN part of the filter, e.g. *wrist watch*, the user will be presented with the information that recovered strings are types of watches, i.e. Grefenstette is exploiting the hyponymy relation inferred by headed compounds. And where *watch* appears in the PRE class, the string is saying something about "things involved in watches", e.g. *watch face*, i.e. the meronymy relation is being exploited. A similar approach is taken in McArthur and Bruza (2000) who use the head-modifier principle to mark a class of automatically returned candidate query refinements to assist the user in query selection.

Grefenstette uses the head-modifier principle to gather together semantically similar terms which are also orthographically related. Ruge (1997) describes how the head-modifier relation can also be used to extract synonymous terms which are orthographically unrelated. Synonyms typically have the same sets of contexts, for example they are modified in the same way: the synonyms *quantity* and *amount* both co-occur with the class of scalar adjectives, and in particular with the scalar adjectives *large* and *small*. The same adjectives repeatedly occurring before two different terms can be used as some sort measure of the two terms' semantic similarity. For compounds, it is the similarity between head elements that is measured. If two heads are semantically similar, there should be an overlap in the modifiers that are found in the separate families of compounds they head. Conversely, orthographically unrelated modifiers can be measured for similarity based on the number of times they share a head.

In the above example the head-modifier principle has been shown to play an important language engineering role, since its value as a principle that relates surface pattern to deeper semantic content has been clearly recognized and exploited.

4 Multi-lingual application of the head-modifier principle in Chinese

The universality of the head-modifier principle in compounding means its application to language engineering can go beyond English. For example, the fact that the dependency of the modifier on the head in the compound is repeated in the paraphrase of the compound has been used to conflate German compound nouns and their phrasal variants (Schmidt-Wigger 1998). Conflation of two structures with the same head-modifier relation, or dependency relation, for automatic indexing of French corpora is presented in Jacquemin and Tzourkemann (1999) and Bourigault and Jacquemin (1999). French is unusual in that it is a left-headed language, yet this in itself does not prevent the construction of head-modifier based filters. An example they give is in (10) and (11) where the phrase structure of (11) is identified with the compound structure of (12).

(11) Noun_1 Prep_2 Noun_3

(12) Noun_1 Noun_3

In this way *fibre de collagene* is related to *fibre collagene* ‘collagene fibre’ where the head in the compound is in left, or first, position which is the same position as in the equivalent phrase. In a collection of papers on the multilingual treatment of nominal compounds, L'Homme (1994) discusses the implications of this difference in linearity of elements in English and French compounds for multilingual applications, and presents a transfer approach to MT where the linearity of Modifier-Head in English is transformed to Head-Modifier in French. Chambers (1994), also looking at English to French MT, is a more detailed analysis of the role of the modifier in an English compound. He characterizes a modifier as one of a range of possible arguments of the Head; the best equivalent in the target language is found by determining exactly which argument it is. Other papers in the same collection make reference to the head-modifier principle for term detection and extraction, as well as machine translation. For example Moreaux (1994) looks at German compound noun detection, and Maalej (1994) describes how the compositionality of English compounds can be exploited for automating English to Arabic translation. Extracting head-modifier pairs for a term, as Strzalkowski et al. (1999) for English, has been done in Spanish for Spanish term conflation (Alonso et al. 2002), where Spanish has left-headed synthetic compounds (e.g. Montrul 1994).

In this section we outline the use of the head-modifier principle for Chinese term extraction. After a few introductory remarks about Chinese compounding, the principal means of lexicon stock expansion in this language, we detail the head-modifier approach to Chinese term extraction from a corpus of Chinese Information Technology texts. The working prototype used is briefly described, including an evaluation.

4.1 Chinese term formation

Chinese belongs to the Sino-Tibetan family of languages which consist of four main groups: Chinese, Miao-Yao, Kam-Thai and Tibeto-Burman (Kratochvil 1968:13). There are seven different Chinese dialects, amongst which are Mandarin, Cantonese and Wu. Though not all are mutually intelligible all dialects use a single writing system such that communication between speech communities is possible through the written word. The unified writing system means that Chinese is the largest linguistic community in the world with over 1.3 billion members (figure from *Ethnologue*). The size of the community makes Chinese a major source of text encoded information requiring extraction methods and techniques. A prerequisite to information extraction that is peculiar to Chinese language texts is a fundamental pre-processing task, namely word segmentation since Chinese natural language texts do not encode word boundaries. Approaches to segmentation have been both symbolic (rule-based), for example Yeh and Lee (1991) and statistical, for example Chen and Liu (1992), Yao and Lua (1998), Peng (2001). Apart from this a major focus of Chinese IE has been the recognition and classification of named entities, a task motivated by the significantly high distribution of proper nouns in newspaper texts. On this, see for example the work reported in Chen and Lee (1996) and Chen, Ding and Tsai (1998) and the National Taiwan University system for proper noun identification described in Chen, Ding, Tsai and Bian (1998).

The vast size of the linguistic community is due to a writing system dating back to at least 1200 BC (Boltz 1996). Chinese is a monosyllabic language where each syllable by default maps onto a morpheme, and morphemes map onto a character in the writing system. For example, the Chinese equivalent of English *multi-media* consists of a string of three morphemes *dūo méi-tĭ*. In the writing system these are represented by the three characters 多媒體 where *dūo*;多 is a free morpheme and *méi-tĭ*;媒體 are bound morphemes, constituting a single free word. Relevant to word structure is the fact that Chinese belongs to the isolating type of languages where the dominant word formation operation is compounding (see Table 1 in section 2). The Chinese equivalents to (1) to (3) in §2.1 are given in (1´), (2´) and (3´)[3]:

(1´)	[電影	協會]
	diàn-yĭng	*xíe-hùi*
	film	society

(2´)	[[電影	協會]	委員會]
	diàn-yĭng	*xíe-hùi*	*wĕi-yuán-hùi*
	film	society	committee

(3´)	[[[電影	協會]	委員會]	醜聞]
	diàn-yĭng	*xíe-hùi*	*wĕi-yuán-hùi*	*chŏu-wén*
	film	society	committee	scandal

When comparing these to the previous examples, what is striking is their structural similarity to English. The head in English is also functionally the head in Chinese: in (1´) *diàn-yĭng* 'film' modifies *xíe-hùi* 'society' in the same way as *film* modifies *society* in the English example. And in (3´) *chŏu-wén* 'scandal' clearly functions as the head as in the equivalent English example. Moreover *chŏu-wén* also determines the syntactic category of the entire structure: *chŏu-wén* is a noun and the compound is a noun. Chinese is clearly headed in that, like English, there is a consistency in the function and location of the head. In other words, what we identify as the head in each compound occupies the same position. More importantly, like English the head is specifically located at the right edge. In others words, Chinese appears to be right-headed. Starosta (1998) presents a convincing argument for right-headed compounds in Chinese, a point acknowledged in Packard's recent (2000) survey of Chinese word structure. Chinese compounds involve elements of all parts of speech, nouns, adjectives, and verbs. The most productive type is noun-noun compounds, as in the examples (1΄) to (3΄). Li and Thompson (1989: 48-54) give a classification of about sixteen sub-types and amongst these there is only one subtype where the head-modifier principle appears not to apply, the so-called *parallel* compound type where neither constitute acts as a head. It should be noted in passing that Huang (1998) argues that Chinese compounds are for the most part not headed but this is because his survey contains many examples of bound morpheme compounds, i.e. where constituents are not themselves words. If it is deemed that 'true' compounds contain constituents that are words, following our definition in section 2, then the assumption is that Chinese compounding is right headed. However there is one major sub-type which appears to be left headed, the so-called resultative verb constructions. For further details, see Li (1990).

Examples of compounds in the vocabulary of information technology are presented in Table 6 and all demonstrate the application of the head-modifier principle. Note that the hyphen denotes bound morphemes which combine to form a word constituent in a compound.

*Table 6. Headed compound terms in Chines*e

Word	Gloss	Modifier	Head
多 媒 體 *dūo méi-tĭ* many media	‘multi-media’	多 *dūo* many	媒 體 *méi-tĭ* media
互 聯 網 *hù-lián wǎng* inter-related net	‘internet’	互 聯 *hù-lián* inter-related	網 *wǎng* net
電 子 郵 件 *diàn-zĭ yóu-jiàn* electronic mail	‘electronic mail’	電 子 *diàn-* electronic	郵 件 *yóu-jiàn* mail

From the examples we see that in each case we have a headed compound, and the head is made up of a free morpheme or two bound morphemes constituting the rightmost element.

4.2 Applying the head-modifier query technique to Chinese term extraction

We have shown how some Chinese compounds are right headed as in English. We can therefore use the same querying method that rests on the head-modifier principle for Chinese as well as English. This is demonstrated with Chinese compound words taken from information technology terminology.

4.2.1 Chinese Information Technology compound terms

In the field of Information Technology, a large number of new terms have to be found to cover a rapidly developing field and many of these have been created by language internal means, in other words with reference to the productive compounding rules of Chinese. Given the arguments above for right-headed compounding in Chinese we would expect newly produced compound terms to be right-headed and therefore subject to our proposed head-modifier query method. Our test data were from a corpus of recently published popular computing articles in a Hong Kong Chinese newspaper *Ming Pao* (specifically the paper's weekly supplement *Hi Tech Weekly*, available at http://www.hitechweekly.com). We collected text published over a six week period (14 June to 24 July 2001), a total of 41364 tokens of Hong Kong Chinese. As an example from the corpus, consider the Chinese word for 'processor', *chǔ-lǐ qì*; 處理器. The structural description is given in (12).

(12) 處理 器 [[chǔ-lǐ]$_V$ qì $_N$]$_N$
chǔ-lǐ *qì*
process tool

The modifier constituent is the root of the compound which is a verb as it is enclosed by internal brackets and labelled with v denoting verb. The entire compound is therefore based on the verb *chǔ-lǐ*; 處理 'to process', the same word used in expressions such as 'to process leather' (Hornby 1999). The head constituent is supplied by the term *qì*; 器 'tool' labelled as a noun. As this is the head the compound term is interpreted to be a type of *tool* which is related to processing. Assuming that the head-modifier principle governs this compound, the constituent *qì*; 器 'tool' can be viewed as a putative hypernym which has a family of hyponyms. We can therefore retrieve its set of hyponyms by a cross-linguistic application of the first search scheme discussed in section 3.1.

4.2.2 Extracting hyponyms within Chinese IT terminology

In (12) the Chinese term *chǔ-lǐ qì*; 處 理 器 ‘processor’ is a right headed compound whose head is *qì*; 器 ‘tool’ and as such can be viewed as one of the set of hyponyms belonging to the term *qì*; 器. Other members of the set will differ only in their modifier element. They can therefore be retrieved by incorporating the head constituent *qì*; 器 into the query used for extracting English hyponym terms [X [substring]]. As the structural description of *chǔ-lǐ qì*; 處 理 器 ‘processor’ shows in (12) a modifier of a right headed compound need not be a noun but in this case is in fact a verb. A search schema similar to the English case is used but the modifier is labelled X_V. This is shown in Table 7.

Table 7. Eliciting hyponyms of qì; 器 *‘tool’*

Information extraction task:	Search schema used:	Query example:
Elicit members of category	[X_V [**substring**]]	[X_V [**qì**]$_N$]$_N$

The results of the query are given in Table 8. As can be seen, the search results are all types of tool whose English equivalents are deverbal agent nouns in *–er*/ *-or*, for example the word for ‘processor’ *chǔ-lǐ qì*; 處 理 器.

Table 8. Elicited hyponyms of qì; 器 *'tool'*

Query string	Elicited strings	English equivalent
[X_V [**qì**]]	1.處 理 器 *chŭ-lĭ* ***qì*** process **tool**	'processor'
	2.散 熱 器 *sàn-rè* ***qì*** scatter-heat **tool**	'cooler'
	3.監 測 器 *jiān-cé* ***qì*** examine-test **tool**	'monitor'
	4.揚 聲 器 *yáng-shēng* ***qì*** raise-sound **tool**	'speaker'
	5.解 碼 器 *jĭe-mă* ***qì*** separate-number **tool**	'decoder'
	6.伺 服 器 *sì-fú* ***qì*** render-service **tool**	'server'
	7.瀏 覽 器 *líu-lăn* ***qì*** swift-skim **tool**	'browser'
	8.掃 描 器 *săo-miáo* ***qì*** sweep-copy **tool**	'scanner'

The results of the query given in Table 8 are all hyponyms of the same term, since the term occupies the head position of the original query. As a next stage we can recast the results of the query as new queries themselves and extend the hyponymy relationship amongst a set of terms. In this case we make reference to the productive noun-noun compounding type in Chinese which is predominantly right headed (see discussion in section 4). One of the targets of the initial query [X$_V$ [**qì**]] is the string *chŭ-lĭ qì* 'processor', the first example in Table 7. We apply the same search schema as before inserting the target string but marking the modifier element as a noun: [X$_N$ [**chŭ-lĭ qì**]].

Table 9. Eliciting hyponyms of the extracted term chŭ-lĭ qì; 處 理 器 *'processor'*

Information extraction task:	Search schema used:	Query example:
Elicit *hyponyms* of extracted term	[X$_N$ [**substring**]]	[X$_N$ [**chŭ-lĭ qì**]]

In this way we target specifically noun-noun compounds whose head is *chŭ-lĭ qì* and which therefore represent the hyponyms of *chŭ-lĭ qì* . The results are given in Table 10.

Table 10. Elicited hyponyms of chŭ-lĭ qì; 處 理 器 *‘processor’*

Query string	Elicited strings	English equivalent
[X$_N$ [**chŭ-lĭ qì**]$_N$]$_N$	1. 伺服器 **處理 器** *sì-fú qì* ***chŭ-lĭ qì*** server **processor**	‘server processor’
	2. 平價 **處理 器** *píng jìa* ***chŭ-lĭ qì*** cheap price **processor**	‘budget processor’
	3. 圖像 **處理 器** *tú-xìang* ***chŭ-lĭ qì*** picture **processor**	‘graphics processor’
	4. 桌面型 **處理 器** *zhūo mìan xĭng* ***chŭ-lĭ qì*** desktop model **processor**	‘desktop processor’

The examples in Table 10 represent the set of hyponyms of a the term *chŭ-lĭ qì*; 處 理 器 ‘processor’, i.e. variety of types of processor, which is exactly what was being queried. The graphical representation of the recursive use of the query schema is given in Figure 5. The mother node represents the hypernym term and the daughter nodes the hyponyms. Daughter nodes can themselves be recast as hypernyms which have hyponym terms, as in the case of *processor* which has *graphics processor*, *desktop processor*, and *server processor* as members, represented as daughters. Each level of the hierarchy is shown with appropriate search schema used to elicit it.

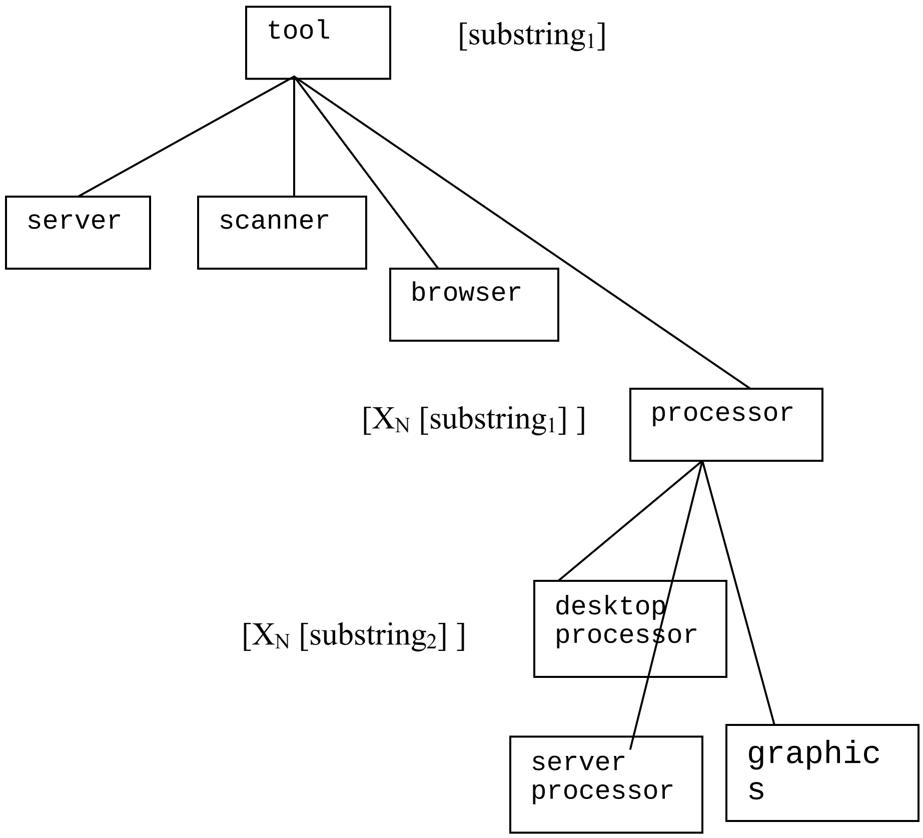

Figure 5. Hyponym hierarchy elicited by recursive application of query [X_N [**substring**]]

4.2.3 Extracting meronyms within Chinese IT terminology

The queries so far have aimed to elicit hyponyms of a given term by assuming the head and querying the modifier. There is another kind of query we can make to elicit attributes of a given member of a given category and so elicit the set of *meronyms* of a key term. The procedure this time is to take the modifier as given and query the head using the second pattern matching schema discussed in §2.2 namely [[**substring**] X_N]. This is shown in Table 11 where what is being queried is the set of terms which constitute the attributes of the term *chŭ-lĭ qì*; 處 理 器 'processor'.

Table 11. Eliciting attributes of the extracted term chŭ-lĭ qì; 處 理 器

Information extraction task:	Search schema used:	Query example:
Elicit attributes of extracted term	[[**substring**] X_N]	[[**chŭ-lĭ qì**] X_N]

It should be carefully noted that the query string itself is identical to that of Table 9 where the query was for hyponyms of the Chinese for 'processor'. The only difference is that in this query the search is for material aligned to the right of the string, i.e. for elements acting as heads in a compound containing *chŭ-lĭ qì*; 處 理 器. Furthermore, the search schema requires the material denoted by X to be tagged as a noun. Again it is noun-noun compounds in Chinese that are most likely to be headed and hence satisfy the information extraction task, in this cases returning terms representing meronyms of the query string. The results are given in Table 12.

Table 12. *Elicited meronyms of* chŭ-lĭ qì; 處 理 器 *'processor'*

Query string	Elicited strings	English equivalent
[[**chŭ-lĭ qì**] X_N]	1. **處理器** 速度 ***chŭ-lĭ qì*** *sù dù* **processor** speed	'processor speed'
	2. **處理器** 型號 ***chŭ-lĭ qì*** *xíng hào* **processor** model	'processor model'
	3. **處理器** 系列 ***chŭ-lĭ qì*** *xì lìe* **processor** series	'processor series'
	4. **處理器** 技術 ***chŭ-lĭ qì*** *jì shù* **processor** technology	'processor technology'

The query results in the table can be viewed as the set of strings which constitute the attributes or properties of the query string *chŭ-lĭ qì*; 處 理 器 'processor', and as such are the set of meronyms of the term. For example, from the results a processor is assumed to have a speed (example 1), a series specification (example 3), a model name (example 2), and so on. In each case what is assumed to be the attribute of the entity is structurally the head of a compound where the entity itself is represented by the modifier element of the same compound.

4.3 Prototype of a multi-lingual Information Extraction System

Given its grounding in a universal principle, the query method outlined applies cross-linguistically. In this section we give a broad overview of a prototype which operates over both English and Chinese texts, focusing on the Chinese component.

4.3.1 Prototype description

The prototype is a distributed system which communicates with the World Wide Web for collecting and pre-processing Chinese language texts, with the Chinese University of Hong Kong's *Jansers* system for word segmentation and part of speech tagging, and with an online Chinese-English dictionary for bi-lingual querying. The system architecture is given in Figure 6.

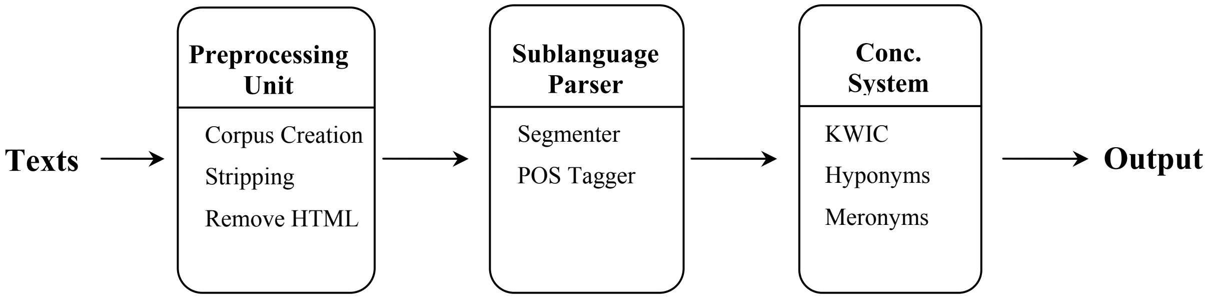

Figure 6. System prototype

The input is html tagged Chinese texts from a newspaper which are stripped and categorised by subject, including Information Technology. Chinese texts do not have explicit token delimiters so a sublanguage parser segments the text and provides part of speech tags for the tokens. Finally a concordancer provides for the presentation of frequency distributions of tokens and their contexts.

The prototype elicits hyponyms and meronyms of key terms using the head-modifier method outlined above. In (13) we have a fragment of a text segmented and tagged.

(13) 接著/DV 再/DV 有/VL 流動/NN 型號/NN 處理器/NN 的/SDG 最/DV 新/A

To elicit hyponyms of the term *qì*; 器 'tool' the search is specified to match all strings with *qì*; 器 'tool' and characters to the left up to the token delimiter. One of the matches will be 處理器/NN 'processing tool, processor' underlined in (13). To then elicit hyponyms of 處理器 the search is this time specified to match 處理器 and the set of characters to the left tagged with NN 'common noun'. One of the matches will be: 流動/NN 型號/NN 處理器/NN 'mobile model processor'. (The system can be made available through prior arrangement with the authors.)

4.3.2 Prototype evaluation

In this section we outline three tests we carried out to measure the performance of the prototype. In each case we used a 41,364 token sample of technical Chinese texts, namely recently published popular computing articles in a Hong Kong Chinese newspaper *Ming Pao*, specifically the paper's weekly supplement *Hi Tech Weekly*, available online at http://www.hitechweekly.com. The first test involved the detection and extraction of nominal compounds, the second and third tests looked at the extraction of hyponymy and meronymy relations between extracted terms.

4.3.2.1 Testing detection and extraction of nominal compounds in Chinese

From Figure 6 we see that an important component of the prototype is the incorporated *Jansers* part of speech (POS) tagger. The output of this component is the input of the compound detection process. Sequences of strings that are tagged as common nouns, i.e. *string*/NN *string*/NN are taken to be headed compounds. We ran the system over the sample of texts and detected 1237 different string sequences tagged in this way. We then looked at each one to determine whether the sequence represented a headed compound noun. The findings are given in Table 13.

Table 13. Compound detection results.

	Type F	Token F	Possible compound	Problematic compound	Score
2 string sequence	941	1384	893	48	94.9%
3 string sequence	245	279	229	16	93.9%
4 string sequence	39	42	38	1	97.4%
5+ string sequence	12	13	11	1	91.2%
Total	**1237**	**1718**	**1171**	**66**	**94.7%**

From the table we see that two string sequences make up the overwhelming majority of string sequences. There were 941 different types of two string sequences (Type F) representing a token frequency of 1384 (Token F). Of these 893 were found to be actual headed Chinese compounds, and 48 were rejected as incomplete or otherwise ungrammatical. This gave a score of about 95% for the prototype's performance of detecting two element compounds. While the two string sequences were the most common, as expected the sequences of five or more strings were the rarest (twelve types in all). Taking all string sequences together, the prototype achieved a score of 94.7% in detecting and extracting headed compounds.

While the results were favourable on the whole it is worth briefly discussing the problematic cases, which make up just over 5% of all the /NN tagged string sequences. In most cases the sequence was found to be incomplete. This was due to the fact that an element tagged with a POS tag other than /NN was omitted. Though most nominal compounds in Chinese are combinations of common nouns, some have verbal or adjectival constituents, and this important group goes undetected. An example is given in (14), which can be glossed as 'hard disk single plate density'.

(14) 硬/A 碟/NN 單片/NN 密度/NN
yìng díe dān pìan mì dù
hard disk single-slice dense-degree
'hard disk single plate density'

Given that our search is based on sequences of the type *string*/NN *string*/NN it is clear that the left-most constituent in (14) will be missed as it is *string*/A. Instead what is detected by the prototype is the incomplete and therefore ungrammatical compound in (15).

(15) 碟/NN 單片/NN 密度/NN
díe dān pìan mì dù

disk single-slice dense-degree
‘?disk single plate density’

To overcome this problem we would need to incorporate queries for nominal compounds whose non-head constituents are existing words belonging to other parts of speech besides common nouns. Grammars such as Li and Thompson (1989) provide a good list of the possible combinations. Another class of problematic cases is where the detected string sequence is an incomplete compound because the left-most constituent is itself part of an already existing compound. Examples of this are given in connection with tests for hyponymy relations which we now turn to.

4.3.2.2 Detecting and extracting hyponym and meronymy relations

To test how the prototype performed in detecting the hyponyms and meronyms of a given core term, we used the Chinese string *chŭ-lĭ qì*; 處理器 ‘processor’ as the representative core term. We ran the prototype over the same sample of texts twice. In the first run we set the string to right-most position to detect all string sequences whose right-most string was 處理器/NN. Extracted strings should therefore be hyponyms of ‘processor’. In the second run we set the same string to left-most position, this time to detect string sequences whose left-most element was 處理器/NN and therefore should constitute meronyms, i.e. attributes, of ‘processor’. The results of these two searches are given in Table 14.

Table 14. Hyponymy and meronymy detection results

	Type F	Possible compound	Problematic compound	Score
				Hyponymy relation
string/NN 處理器/NN	13	11	2	84.6%
				Meronymy relation
處理器/NN *string*/NN	16	16	0	100%

From the table we see in the sample of texts there were thirteen different string sequences with 處理器/NN appearing in right-most position, eleven of which on inspection were viewed as hyponyms of ‘processor’, giving the prototype a score of 84.6%. For meronyms sixteen different string sequences were detected, this time with 處理器/NN on the left, and all of these we analysed as attributes of ‘processor’. It is worth looking briefly at the problematic cases for the hyponymy test.

One of the extracted string sequences is given in (16). We clearly cannot interpret (16) as a hyponym of ‘processor’: there is no sense in which ‘model processor’ is a type of processor.

(16) 型號/NN 處理器/NN
xǐng hào chǔ-lǐ qì
model processor
‘?model processor’

The problem lies in the fact that the left constituent is itself part of an already existing compound, one of whose constituents has been ‘missed’. This compound is ‘mobile model’, as shown in (17).

(17) 流動/NN 型號/NN
lĭu dòng xĭng hào
mobile model
‘mobile model’

This already existing headed compound can freely combine with ‘processor’ to act as a complex non-head constituent of a three element compound. For clarity we give the constituent structure in (18) showing the head ‘processor’ as attaching to an already existing compound ‘mobile model’. The three sequence string representing this three element was in fact detected by the prototype and is given in (19).

(18) [[mobile [model]] processor]]]

(19) 流動/NN 型號/NN 處理器/NN
lĭu dòng xĭng hào   chŭ-lĭ qì
mobile model processor
‘mobile model processor’

Another example of the same kind is the Chinese for ‘test version processor’. In this case left constituents are part of the already existing compound ‘test version’. However the prototype extracts the string sequence in (20) where the modifier of the already existing compound ‘test’ has been omitted yielding the incomplete ‘?version processor’. The complete compound is also detected and is given in (21).

(20) 版本/NN 處理器/NN
băn-bĕn chŭ-lĭ qì
version processor
‘?version processor’

(21) 測試/NN 版本/NN 處理器/NN
cè-shì *bǎn-běn* *chǔ-lǐ qì*
test version processor
‘test version processor’

The above examples show where the prototype has overgenerated by supplying string sequences which are not hyponyms of the core term. We also discovered examples where the prototype has undergenerated by failing to detect a hyponym that exists in the sample text. One interesting case concerns the use of mixed fonts in Chinese texts. In specialist texts the terms can originate conceptually from a non-Chinese source, typically English. In such cases the English term may be borrowed together with its orthography. The POS tagger fails to tag any string which is not in Chinese characters. If the string happens to be a non-head constituent of a compound, as an untagged string it will not be detected by the prototype. In (22) we see that the sample contained the alternate compound for ‘mobile model processor’ where the equivalent for ‘mobile’ is a direct orthographic borrowing from the English.

(22) Mobile 型號/NN 處理 器/NN
xǐng hào *chǔ-lǐ qì*
mobile model processor
‘mobile model processor’

However, since the ‘mobile’ constituent is untagged it goes undetected and hence the compound itself in is not detected.

5 Concluding remarks

We began with the observation that while nlp/e has been a boon to some areas of linguistics, equally linguistic theory can be used to enhance methods and techniques in nlp/e. This is because the disciplines are connected by a common object of enquiry, natural language. In this context we have discussed a theoretically driven method for language engineering where the linguistic insight is the head-modifier principle, a universal constraint on the structure of words. We have shown how the method can be used for extracting sets of multi-word terms in a document that are defined with reference to the *type hierarchy* that is central to ontology, where the relationship is governed either by hyponymy or meronymy (see Sowa 2000: 492-494). The method has a depth of application, given that the majority of terms are multi-word. Because of the language universal principle on which it is based nature, the method also has breadth of application: on the one hand it is relevant to any subject domain that is describable through natural language texts; and on the other it can be applied cross-linguistically, as we have shown through a case study of Chinese IT term extraction

In another sense the method has a fairly restricted application given the presuppositions bound into it. First, it presupposes compound terms. Of course not all new terms are compounds. They may be direct borrowings, or created by means of another word formation operation such as affixation. Second, it presupposes headed compound terms. But not all compounds are headed, and even amongst the noun-noun compounds where headedness dominates it is possible to have exocentric, or unheaded, compounds. Third, it presupposes heads to be the right most element. This is the case in English and Chinese, but some languages have left-headed compounds, such as the

Romance languages, as we mentioned in our discussion of French and Spanish compounding and nlp in §4.

Nonetheless compounding is a highly productive way of coining new terms, particularly in special languages, and the major class of compounds in a language are headed, and for most languages the head is the right most constituent. Given this we have outlined a potentially powerful multi-lingual term extraction method that searches through a range of language documents and semi-automatically organises ontological type hierarchies amongst key terms thus capturing some of the information structures present yet implicit. Integral to the method are theoretical claims about the linguistic properties of the terms themselves and as such it represents the ways in which insight from language theory can be profitably employed for the benefit of language engineering.

References

Abney, S. 1991. Parsing by chunks. In: R. C. Berwick, A. Abney and C. Tenny (eds) *Principle-Based Parsing: Computation and Psycholinguistics*. Dordrecht: Kluwer. 257-278.

Alonso, M. A., Vilares, J. and Darriba, M. 2002. On the usefulness of extracting syntactic dependencies for text indexing. In: O'Neill, M., Sutcliffe, F., Ryan, C. and Eaton, M. (eds) *Artificial Intelligence and Cognitive Science*. Berlin: Springer Verlag. 3-11.

Anderson, S. R. 1985. Typological Distinctions in Word Formation. In: T. Shopen (ed.). *Language Typology and Syntactic Description, vol. III*: *Grammatical Categories and the Lexicon.* Cambridge: Cambridge University Press.30-56.

Bauer, L. 1994. Head and Modifier. In: R. E. Asher (ed.) *The Encyclopedia of Language and Linguistics* vol. III. Oxford: Pergamon Press.1529-1532

Beard, R. 1998. Derivation. In: Andrew Spencer and Arnold Zwicky (eds) *The Handbook of Morphology*. Oxford: Blackwell. 44-65.

Boguraev, B. and Pustejovsky, J. 1996. *Corpus Processing for Lexical Acquisition*. Cambridge, Ma.: MIT Press.

Boltz, W. G. 1996. Early Chinese Writing. In: Daniels, P. and Bright, W. (eds) *The World's Writing Systems*. Oxford: Oxford University Press. 191-199.

Bouillon, P. and Estival, D. (eds) 1994. *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*. Geneva: ISSCO.

Bourigault, D. and Jacquemin, C. 1999. Term extraction and term clustering. In *Proceedings of the European Association of Computational Linguistics*. Bergen. 15-22.

Chambers, C. 1994. Analysing and Generating English Compound Structures for Machine Translation. In Bouillon, P. and Estival, D. (eds) 1994. 125-134.

Chen, H. H. and Lee, J. C. 1996. Identification and Classification of Proper Nouns in Chinese Texts. In *Proceedings of 16th International Conference on Computational Linguistics*, Copenhagen, Denmark. 222-229.

Chen, H. H., Ding, Y. W. and Tsai, C. T. 1998. Named Entity Extraction for Information Retrieval. *Computer Processing of Oriental Languages* 12 (1), 75 – 85.

Chen, H. H., Ding, Y. W., Tsai, S. C. Bian, G. W. 1998. Description of The NTU System Used for MET2. In *Proceedings of the 7th Message Understanding Conference* Washington DC. [published at http://trec_nist.gov/pubs/trec10/t10_proceedings.html]

Chen, K. J. and Liu, S. H. 1992. Word Identification for Mandarin Chinese Sentences. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, Nantes. 101 - 107.

Chomsky, N. and Halle, M. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Corbett, G. G. and Fraser, N. 1993. Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics* 29, 113-142.

Evans, A., Ginther-Webster, M., Lefferts, R. and Monarch, A. 1991. Aotomatic Indexing Using Selective NLP and First-Order Thesauri. In: Lichnerowicz, A (ed.) *Intelligent Text and Image Handlings*. London: Elsevier. 624-643.

Evans, R. and Gazdar, G. 1996. DATR. *Computational Linguistics* 22 (2), 167-216.

Felber, H. 1984. *Terminology Manual*. Paris: UNESLO International Information Centre for Terminology.

Frantzi, K. T. and Ananiadou, S. 1997. Automatic term recognition using contextual clues. In *Proceedings of Mulsaic* 97, IJCAI. Japan.

Fraser N. M., Corbett, G. G. and McGlashan, S. 1993. *Heads in Grammatical Theory*. Cambridge: Cambridge University Press.

Grefenstette, G. 1994. *Exploration in Automatic Thesaurus Discovery*. Kluwer: Academic Publisher.

Grefenstette, G. 1997. SQLET: Short query linguistic expansion techniques, palliating one-word queries by providing intermediate structure to text. In *Proceedings of the RIAO '97*. Montreal. 500-509.

Hacken, P. 1995. The Role of Pronominal Reference in Compound Detection. In: Bouillon, P. and Estival, D. (eds) 1995. 1-14.

Haugen, E. 1950. The analysis of linguistic borrowing. *Language* 26, 210-31.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*. Nantes. 539-545.

Heid, U. 1999. Extracting Terminologically Relevant Collocations from German Technical Texts. In: Sandrini, P. (ed.). *Terminology and Knowledge Engineering, Proceedings of the 5th International Conference*. Vienna. 23-27

Hippisley, A. 2001. Word Formation Rules in a default inheritance framework. In: Jaap van Marle and Geert Booij (eds) *Yearbook of Morphology 1999*. Dordrecht: Kluwer.221-261.

Hornby, A. S. 1999. *Oxford Advanced Learner's English-Chinese Dictionary*, (Extended 4th Edition). Hong Kong: Oxford University Press.

Huang, S. F. 1998. Chinese as a headless Language in Compounding Morphology. In: Packard, Jerome L. 1998. 261-83.

Jacquemin, C. and Tzourkemann, E. 1999. NLP for term variant extraction: synergy between morphology, lexicon and syntax. In: Tomek Strzalkowski (ed.)1999. 25-74.

Justeson, J. and Katz, S. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1 (1). 9-27.

Kratochvil, P. 1968. *The Chinese Language Today*. London: Hutchinson & Co.

L'Homme, M. 1995. Traitement des groupes nominaux en traduction automatique. In: Bouillon, P. and Estival, D. (eds) 1995. 147-161.

Lauer, M. 1995. Corpus statistics meet the compound: some empirical results. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. 47-55.

Li, C. N. and Thompson, S. A. (1989) *Mandarin Chinese: a functional reference grammar*. Berkeley: University of California Press.

Lieberman, M. and Sproat, R.. 1992. The stress and structure of modified noun phrases in English. In: Ivan Sag and Anna Szabolcsi (eds) *Lexical Matters*. Stanford: CSLI.131-181.

Maalej, Z. 1995. English-Arabic Machine Translation of Nominal Compounds. In: Bouillon, P. and Estival, D. (eds) 1995. 135-146.

McArthur, R and Bruza, P. 2000. The Ranking of Query Refinements of Interactive We-based Retrieval. In *Proceedings of the Information Doors Workshop*.

Montrul, S. 1995. Headedness and Argument Structure in Spanish Synthetic Compounds. In: Bouillon, P. and Estival, D. (eds) 1995. 44-60.

Moreaux, M. 1995. Détection, segmentation et interpretation des noms multi-lexicaux allemands. In: Bouillon, P. and Estival, D. (eds) 1995. 88-102.

Packard, J. L. (ed.) 1998. *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*. Berlin / New York: Mouton de Gruyter.

Packard, J. L. 2000. *The Morphology of Chinese- A Linguistic and Cognitive Approach*. Cambridge: Cambridge University Press.

Peng, F. and Schuurmans, D. 2001. Self-supervised Chinese Word Segmentation. In F. Hoffman et al. (ed.) *Advances in Intelligent Data Analysis*, *Proceedings of the 4th International Conference (IDA-01)*. Heidelberg: Springer-Verlag. 238 – 247.

Pustejovsky, J. Bergler, S. and Anick, P. 1993. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics* 19 (2). 331-358.

Riloff, E. and Shepherd, J. 1999. A Corpus-base Bootstrapping Algorithm for Semi-Automated Semantic Lexicon Construction. *Natural Language Engineering* 5 (2). 147-156.

Rogers, M. 1997. Synonymy and equivalence in special-language texts. A case study in German and English texts on Genetic Engineering. In: A. Trosborg (ed.). *Text Typology and Translation*. Amsterdam/Philadelphia: John Benjamins.217-245

Ruge, G. 1997. Automatic detection of thesaural relations for information retrieval applications. In: *Foundations of Computer Science: Potential - Theory – Cognition*. Heidelberg: Springer. 499-506.

Sager, J. C., Dungworth, D. and McDonald, P. F. 1980. *English Special Languages.* Wiesbaden: Oscar Brandstetter Verlag.

Salton, G. and McGill, M.J. 1983. *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Schmidt-Wigger, A. 1998. Building Consistent Terminologies. In *Proceedings of COMPUTERM*. Montreal.

Soderland, S., Fisher, D., Aseltine, J. and Lehnert, W. 1995. CRYSTAL: Inducing a Conceptual Dictionary. In *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*. 1314-1319

Sowa, J. F.2000. *Knowledge Representation*. London/NewYork: Brooks-Cole.

Sparck Jones, Karen. 1985. Compound noun interpretation problems. In: Frank Fallside and William Woods (eds*) Computer Speech Processing*. London: Prentice Hall. 363-381.

Spencer, A. 1991. *Morphological Theory: an Introduction to Word Structure in Generative Grammar*. Oxford: Blackwell.

Starosta, S. et al. 1998. On defining the Chinese compound word: headedness in Chinese compounding and Chinese VR compounds. In: Packard, J. 1998.347 – 370.

Strzalkowski, T. (ed.). 1999. *Natural Language Information Retrieval*. Dordrecht: Kluwer.

Williams, E. (1981). On the notions 'lexically related' and 'head of a word'. *Linguistic Inquiry* 12, 245-74.

Yao, Y. and Lua, K. T. 1998. Splitting-merging Model of Chinese Word Tokenization and Segmentation. *Natural Language Engineering* 5, 309 – 324.

Yeh, C. L. and Lee, H. J. 1991. Rule-based Word Identification for Mandarin Chinese Sentences: a Unification Approach. *Computer Processing of Chinese & Oriental Languages*, 5 (2): 171 - 184.

Zwicky, A. M. 1985. Heads. *Journal of Linguistics* 21, 1 – 29.

Zwicky, A. M. 1993. Heads, bases and functors. In: Fraser, Corbett and McGlashan 1993. 292-315.

Ethnologue: http://www.ethnologue.com/web.asp

Jansers: http://www.jansers.org/new/big5/service.html

Notes

[1] The authors are grateful for the support from the EU IST project *Gida* (grant # 2000-31123) and the EPSRC project *Socis* (grant # GR/M89041). We would also like to thank the anonymous referees for their helpful comments. We are grateful to one referee who pointed us to a wealth of literature on the use of the head-modifier relation in nlp.

[2] See Hacken (1995) for a detailed discussion on other criteria used to define compounds. One of these is the blocking of pronominal reference to the left-headed element in a compound, which is also used as evidence of its wordhood. Lieber and Sproat (1992) give a detailed X-bar approach to distinguish 'true' compounds which are lexical object, from phrasal categories, which are syntactic objects.

[3] The examples in (1´) to (3´) have been tested by two Chinese native speakers, one from Beijing, China and the other from Hong Kong Special Administrative Region, China..