AUTOMATED CHINESE TERMINOLOGY BUILDER

Suen Caesar LUN

City University of Hong Kong

1. Introduction

Descriptive linguists usually hold very negative ideas about standardization of language use. They believe the role of linguistics is simply, which already amounts to a great deal of work, to discover the patterns and distribution of language use, the internal linguistic system within the brain, and the historical and sociolinguistic changes that occur in different components of language, etc. The role they want to play is an observer, a critical analyst, and a scientist above all. However, if the knowledge of linguistics is not utilized to facilitate language-related tasks, such as language teaching, translation, lexicography, and terminology, it is not just a waste of knowledge, but a mistake of blindfolding ourselves when choosing which way to go. Nevertheless, it is not to say that the task of terminology is only a task for linguists. In fact, terminology is an interdisciplinary job which requires dedicated efforts from experts from all fields, including linguists, lexicographers, lexicologists, computer scientists, cognitive psychologists and specialists of each subject field, and possibly also many other fields too.

Terminology can be standardized in different ways. First, different alternatives can be invented by people involved, such as specialists in the field, or more likely, by translators. After a certain period of contention, one of the alternatives will be adopted or different alternatives may find different usage, thus co-exist. Second, a national institution for standardization may collect opinions and finally determine which term to be codified as standard usage. Third, deliberate and systematized coinage of terminology by a panel of concerned parties. The three ways mentioned above need not exclude each other but each has its pros and cons.

The first method is probably the most common one and it requires not much coordination. However, it takes time and therefore will slow down the standardization process. The terms thus coined may show a high degree of arbitrariness. This is a bottom-up approach and is rather democratic. The second method probably requires a lot of coordination and centralized efforts, and is therefore very expensive. The terms thus coined need some channel for specialists in the field to recognize and adopt. Moreover, the feeling that many specialists are left out of the decision process may create some obstacles in the smooth adoption. This is a top-down approach and is quite authoritative in nature. The third method is to pull the strengths of all those involved into the business and let the computer do some preliminary work to provide alternative terms according to linguistic findings. The alternative terms thus coined can then be studied by specialists in the field/s and finally the government can step in and publicize the consensus. This paper will propose an automated Chinese terminology builder and discuss how it can serve as an integral part of the third method.

2. The Chinese language situation

When one mentions the terminology issue in the context of the Chinese language, the situation is rather complex. People use different vernacular dialects in the oral speech though we all claim to share the same written language. It is even more complicated because of the existence of two competitive political entities, namely the PRC government and the government in Taiwan, not to mention other smaller Chinese communities such as Singapore, Hong Kong Special Administrative Region and Macau. So if you ask any Chinese what authority they turn to when there is a question about standardization of terminology, one will not find an easy answer. Even for people with more knowledge, they are not sure whether the Language Commission or the Academy of Social Sciences is responsible. Another complication is the script. One character can be written in different ways depending on whether the simplified script or the traditional script is used. Even when one script is agreed upon, there is still argument concerning which character form to use, which is mainly due to the debate of reversal to the original form or adoption of a commonly accepted form, the so-called "similar morpheme with different graphemes" issue.

However, there are certain advantages in the Chinese language when dealing with word formation. First, Chinese people are familiar with the morphemes of the language. Unlike European languages, where morphemes can be inherited from classical languages such as Latin and Greek, which may not be very familiar to the ordinary people, Chinese morphemes are taught as characters (hanzi) since the kindergarten and the morpheme senses are therefore not unfamiliar to them. A graduate from the primary education in Hong Kong has to master 2,600 characters (each character may represent more than one morpheme) Second, the written form does not change as radically as English morphemes, due to a lack of phono-morphemics, it is easier for people to associate character form with morphemic sense. Third, the word formation principle is rather straightforward and structures at different levels can be described as making use of the same principles. More will be described in the following section. In short, this paper wants to propose a theoretical model of automated terminology coinage based on Chinese word-formation principle that can bypass political controversies, as well as one that can be implemented mechanically and systematically.

2.1. Chinese word formation principles

Chinese can make words by means of different components of language. Take the phonological component for example, there may be coinage, onomatopoeia, repetition of rhymes or alliteration, and transliteration.

Coinage: mao1 (cat), gou3 (dog)

Onomatopoeia: bu4gu3 (cuckoo), ding1ling1dang1lang1 (dingdong)

Reduplication:

Whole morpheme: piao1piao1ran2 (complacent, self-satisfied), pian1pian1- qi3wu3 (dance lightly)

Repetition of rhymes: hun4dun4 (Chaos), hun2tun1 (wonton)

Alliteration: gan1ga4 (embarrassed), jie2ju1 (in straitened circumstances, short in money)

Transliteration of loan words: sha1fa (sofa), ji1pu3che1 (jeep)

The second class is to use the morphological means of affixation. However, the number of affixes is small in Chinese.

Prefixation: lao3-hu3 (tiger), lao3-shu3 (rat)

Suffixation: hai2-zi (child), xie2-zi (shoe), hua1-r (flower), hua4-r (painting)

The third class is syntactic in nature, or so-called compounding. Most of the words are formed by means of combining two root morphemes or more together according to the same principles which are used in forming phrases and clauses. This is one of the reasons why computerized segmentation of a string of Chinese text into separate words could be quite difficult because the word boundary and phrasal boundary is sometimes hard to tell.

Compounding:

SP (subject-predicate): di4zhen4 (earthquake), xin1-tong4 (heartache, sad)

MH (modifier-head): huo3che1 (train), bai2cai4 (bok choy), xia1mang2 (be busy for nothing)

VO (verb-object) : da3zi4 (typing), chuan2zhen1 (fax), shuo1hua4 (talk)

VC (verb-complement): gai3shan4 (improve), tiao4gao1 (high-jump), jiu1zheng4 (correct, rectify)

CONJ (conjoined (synonymous or anonymous)): peng2you3 (friend, companion), shi1sheng1 (teacher and pupil), da4xiao3 (size)

There is also a method that makes use of measure words to form collective nouns. Since measure words are non-existent in English, this is therefore a unique Chinese way, e.g. niu2zhi1 (cattle), ma3pi3 (horses), ren2kou3 (population), shu1ben3 (books).

Abbreviation:

Abbreviation means merging words together to form single words by clipping the most salient morphemes of each word in the compound. The resultant word still has to comply with the rules for compounding, e.g. Bei3jing1 Da4xue2 (north capital big school) > Bei3da4 and Di4yi1 Zhong1xue2 (number one middle school) becomes Yi1zhong1. So saliency means the degree of how the morpheme can help distinguish a word from another. So the generic morpheme for ‘school’ xue2 is not used in the abbreviation process. Moreover, the resultant words still keep a modifier-head structure.

Among these methods, compounding and abbreviation are most frequently employed. In forming new terms, we rely on the existing morphemes/characters, and words. Therefore, coinage in the phonological sense is usually not a good method. That means, to make the terms more intuitive and therefore much easier to associate and memorize, Chinese word formation principles must be observed. And among these principles, since we are not making core vocabulary, compounding may be the most productive method to adopt. However, if the terms are too long but the concepts are very familiar to specialists in the field, there is a tendency for them to be abbreviated. So we can view compounding and abbreviation as two dynamic means in naming concepts. One acts like a generator and the other acts like an equalizer, making the most economical use of resources in writing and speaking. Next, let us turn to the morphemic analysis.

2.2. Chinese morphemic analysis

Although the morphemic senses of hanzi are relatively transparent to most Chinese, the knowledge of morphemics is not. There is not a one-to-one correspondence between a hanzi and a morpheme. Usually a hanzi may represent one or more morphemes, and the morphemes represented by the same hanzi may or may not be historically related. For instance, ben3 could mean either 'tree trunk' or 'origin' and kou3 either 'mouth', or 'measure word for eating, or animal', but hui4 either could mean ‘will’, or ‘meeting’, which are not related. Almost all native Chinese morphemes are monosyllabic and the multi-syllabic morphemes were basically borrowed from other languages. So, only in terms of Chinese morphemes is it correct to claim that the Chinese language is monosyllabic. Chinese words are not monosyllabic in nature in the modern sense.

To coin terminology in any field, we need to observe the Chinese word formation principles and we need to know more about the morpheme senses, their distribution and frequency. At Tsinghua University, Beijing, Yuan et al. (1994) developed a morpheme database that tries to describe all Chinese morphemes for the purpose of semantic tagging and unknown word processing of running texts. Each morpheme will be dealt with in a worksheet which records 10 attributes, namely, morpheme, sense number, pronunciation key, sense description, category, word-formation degree, position, semantic features, prosodic features, remarks. Moreover, all the words collected from dictionaries will be tagged with such morphemic information so that we can tell how many words are formed according to which principle, which morphemes are core and productive as opposed to the less frequently used ones, and when the morphemes are used, whether they are used as bound or free morphemes, and whether they are used before or after another morpheme. Once such information is gathered, its application should guarantee a more robust segmentation algorithm.

This morpheme database should be very valuable for terminology purposes and it will be even nicer if semantic properties are also considered, such as hyponymy, antonymy and synonymy, further sub-categorization of verbal morphemes, such as transitivity, reflectivity, subject animacy, object animacy etc. With an up and running morpheme database, the other properties can be generated with much ease. However, to accurately reflect what has been newly injected into the language as new words or existing words with new usage, running corpora should be used to check any new changes. The latter task, of course, changes the static description into a dynamic description of actual language use. However, it will be a very expensive job that is hardly affordable by any short-term research projects. Now, let me turn to Alshawi's work on Longman Dictionary of Contemporary English (LDOCE).

3. Parsing of sense definition

Parsing means to extract a hierarchical structure from a linear string of symbols. For LDOCE, there is a restricted definition vocabulary of around 2,000 words. With such a vocabulary, all the sense definitions can be described. It is said that this vocabulary has more in common with a “Basic English” vocabulary than a set of semantic primitives. Alshawi's project is to extract information for semantic classification of words in LDOCE by parsing the sense definitions. It is most feasible when the sense definitions are written in a systematic way, possibly according to some policy paper as described by Sinclair (1987). Please see the following example from Alshawi for the noun 'launch' (Alshawi 1989):

```
(launch)
(a large usu. motor-driven boat used for
 carrying people on rivers, lakes, harbors, etc.)
((CLASS BOAT) (PROPERTIES (LARGE))
      (PURPOSE
             (PREDICATION (CLASS CARRY) (OBJECT PEOPLE))))
```

As one can see, during parsing, specific structures are used to indicate different information. The head of the initial noun phrase for a noun entry usually contains the genus or superordinate information with or without modification, key words such as 'used for' indicates purpose and the remainder indicates further differentiae that differentiate this word from other semantically related words. It should be noted that different parts of speech may be defined with different formats, so a POS specific set of phrase structure rules may be invoked to facilitate parsing. So, provided with a dictionary in good lexicographical style, a lot of useful semantic and collocational information can be extracted for each headword. The result may be a natural (in the sense that it is not based on semantic primitives conjectured by the linguists but using natural language as 'meta-language') semantic network which can be used in different applications. There are shortcomings in Alshawi's work according to himself, but the general picture is clear for our purpose.

4. Requirements for Terminology

Terminology refers to a set of terms used in a particular science or discipline. It is therefore universal in nature as opposed to ordinary words in a national language which are more culturally loaded. However, since Chinese is so different from English, transliteration, which simply changes one spelling system into another, such as changing the Roman alphabet to the Cyrillic alphabet or Greek alphabet, cannot be done easily as among European languages. The reason is that a syllable can be represented by more than one hanzi, thus no uniformity is found. Even the sounds can be represented by the same hanzi, the meanings are not transparent, e.g. 'bus' is ba1shi4 but ba1 means ‘place’ while shi4 means ‘scholar’, none has any relation to the concept of 'bus'. Therefore, loan translation, or 'calque' is more prevalent as an alternative. National pride or language planning may not be a major reason in accounting for the trend to replace transliterations with loan translations as time passes by. Self-adjustment for the purpose of higher transparency in meaning and better communication should be the dominant driving forces.

It should be noted that terminology is different from an ordinary dictionary in that we are more concerned with the concepts rather than the linguistic patterns of the terms. This, however, does not mean that the latter is of less importance. The terms in a particular terminology should be coherently coined and should reflect the state of art of that field as well as the internal relationships among the concepts. Consequently, it is important to be able to find a structure out of the terms and different terms should be inter-related through different semantic relationships or attributes. For example, we like to know in a particular field, the key concepts, the compartmentation of the disciplines, the objects, processes, events involved and how, the whole-part relationship, the cause and effect, action and reaction and various kinds of relationship that can relate one term to another. We can even think of the terms as key players in a script, or frame, or scenarios as knowledge system experts may like to call them. Knowing the terms well should imply an understanding of what is going on in the field.

The purpose of automatically generating alternatives for new terms is that we want to achieve consistency in the nomenclature, and thus resulting in transparency within the terminology. The results will not only facilitate our understanding of the concepts, but also enable us to recall the terms more effectively. Now, let me try to elaborate on the system.

5. System Architecture of Termcoiner

5.1. Knowledge Bases

The cylinders in Appendix A represent knowledge bases that the system needs in order to coin terms. They are:

a. Corpus of Definitions (computer terminology for example);

b. Rule Base;

c. Lexicon;

d. Morpheme Database;

e. Dynamic Statistical Rules of Chinese Word Formation Processes;

f. Character Set;

h. Database of Core and Translated Terms.

The basic assumption is that there are ideas technical dictionaries in English, the universal language for sciences that have well written sense definitions. In reality, that may be far from the truth. So, complexity of parsing depends a lot on the language structures found in the technical definitions in the corpora. Here we have taken Webster' dictionary and Microsoft's dictionary for computer science as sample definitions. The style of defining the terms is really quite dissimilar in the two. The rule base consists of phrase structure rules for parsing the definitions. Subsets of rules can be invoked to parse different parts of speech to speed up the process. The lexicon means a combination of ordinary vocabulary and technical dictionary in the field concerned. The morpheme database, as mentioned above, consists of information that enables us to choose the right morphemes for a new coinage.

Dynamic statistical rules consist of frequency counts of different kinds of POS and may vary in different fields. For instance, for describing a machine or a tool, it is highly plausible that the word is a tri-morphemic word of the structure ((van)VO, n)MH or ((v,v)CONJ, n)MH, where v, n, mean verbal morpheme and nominal morpheme and MH, CONJ represent compounding according to Section 2.1. So, a punching machine is a da3 (type/strike) kong3 (hole) ji1 (machine), and a printer is a da3 (type/strike) yin4 (print) ji1 (machine). Nouns and verbs may use different word formation principles to a different degree. By analyzing dynamically and empirically the terms in technical dictionaries, various kinds of information of existing terms can be discovered which serve as statistical rules to guide the coining process. The character set is derived from the keywords obtained from the parse trees of the definitions. Using information from the morpheme database, it provides a list of possible synonymous morphemes, e.g. ji1 (machine), qi4 (device, machine), ju4 (ware, tool) etc. English examples might be dio-, theo- or god, or love, ami-, amor, phil- etc. The purpose of having this knowledge is that different morphemic forms may have different properties, and their productivity in word coinage may be different. Another consideration is that homonyms can be avoided in order for the terms to be more distinguishable. Here, the question is whether we need to take phonology into consideration. For example, Chinese does not have verbs which are four characters long, except in classical idioms used as verbs. The database of core and translated computer terms is derived from the coinage of the first generation. The core terms then serve as seminal terms that can be directly employed in coining compound terms, e.g. 'database management' can simply be rendered into 'zi1liao4ku4 + guan3li3' once the two words have been separately coined and stored. Of course, there is still a possibility that an abbreviation is desired. So the long form and the abbreviation become two alternatives for terminologists and experts to decide and both may be kept.

5.2. The Flow

The input can be either a simple word or a compound word. If it is a simple word, we may still use the tranliteration strategy as in 'bit' (translated as bi3te4). So we need to get the IPA transcription and use the Syllable-based Transformer to come up with a transliteration. Of course, the preferred method is to have a loan translation. In fact, it is more common now to call ‘bit’ as wei4yuan2, meaning ‘position’ + ‘element’. So following that path, one has to acquire the definition/s of the word from the Corpus of Definitions and pass it/them to the Parser. The parse tree/s generated will be analyzed and key words will be extracted. Key words may be headwords in syntactic constituents and their relationships can be determined by the grammatical structure or function words. Then with the help of the Character Set, the Morpheme Database, and the Dynamic Statistical Rules, various alternatives will be generated.

Core words thus coined will be scrutinized by terminologists and experts and then stored in the Database for generating compound words.

If the input is a compound word, it will be broken into individual words. If any of these words is not already in the Core Database, it can be passed to the simple word coiner. When all the words are translated, sometimes in a one-to-one manner, they will be grouped together and an abbreviation may be sought further. The output will be alternative Chinese equivalents for an English term.

5.3. Some examples

Here are dry runs of some authentic examples:

6. Initializing means formatting a disk.
ge2shi4hua4 ci2die2

8. Login means signing in on a computer.
qian1ming2 ru4/jin4

9. An interpreter is a program that performs interpretation.
cheng2xu4 jin4xing2 jie3shi4

16. An instruction is usually made up of an operation code and one or more operands.
cao1zuo4 ma3 cao1zuo4shu4

21. An instruction is a group of characters, bytes, or bits, that defines an operation to be performed by the computer.
zi4 zi4jie2 zi4yuan2 ding4yi4 cao1zuo4

And down below are the Chinese equivalent terms found in a Mainland Chinese computer Dictionary:

initialize = yu4zhi4/chu1shi3hua4
login = jin4ru4 (enter into a system), qian4dao4, gua4hao4,zhu4ce4
interpreter = jie3shi4ji1/jieshi4 cheng2xu4
instruction = zhi3ling4

It is assumed that the underlined words are the keywords found after definition parsing following Alshawi' approach. The Chinese equivalents are taken from the Lexicon. By following the flow, 'login can have 'qian4dao4 as a possible alternative generated by the system, similar to 'qian1ru'. The difference is that dao4 means 'at' and ru4 means 'in'. 'Interpreter' should be alright because it is a compound of jie3shi4 and cheng2xu4. Whether cheng2xu4 is synonymous to ji1 (machine) depends on whether there is such a registration in the database. For 'initialize', and instruction, the definitions do not provide too much help, but it does not mean that the equivalent of 'disk-format' cannot be taken as a proper translation, hence ci2die2ge2shi4hua4. We quote two definitions for instruction here to show how diverse definitions can be. From the first instance, we may get cao1zuo4shu4ma3 (operation element and code); whereas from the second instance, cao1zuo4ding4yi4. Both are acceptable but not as brief as direct translation of the term in ordinary Chinese zhi3ling4. This is true of the term 'to initialize'.

So, direct translations may be a source of alternative terms, transliteration too. The point driven here is that since Chinese morphemes are transparent in meaning and word formation principles concentric with other larger syntactic structures, a logical and sensible way is to provide grammatically well-formed alternatives for terminologists and experts to examine. Cases like calling a mouse of a computer a mouse may not be a very productive method in the sense that a Chinese mouse may not be as cute a translation as Mickey Mouse in the Western tradition. For a scientific term, it should be clearer if the translation is determined by its functions and class. So a mouse should better be called a pointing device or an electronic device for moving the cursor around, thus zhi3biao1qi4 or yi2biao1qi4. Other synonymous morphemes can be used. However, judging from the attributes of these morphemes, the term finally arrived at should not be too far away from the suggestions.

6. Conclusion

The terminology of any field is not totally blank. It is impossible to undo all the terms and coin them from scratch again. However, for those existing ones that are not very transparent or those that have different ways to represent due to the separation of the Chinese communities, re-consideration is not at all impossible. It is hoped that a terminology should be consistently created following the principle of 'linguistic integrity'. The terms should not be randomly coined, with no checking with other related words in the field. The terms should be formed according to conventional Chinese word formation principles. They had better be homonym free as far as possible. They should also be communicatively effective, for both learners and users.

The prerequisite for Termcoiner to be successful is to have good technical dictionaries where definitions are written up carefully in a consistent style. This may be very demanding and not realistic at present. However, if we consider the work of Sinclair (1987), we can actually design worksheets for experts to fill in attributes which answer all we need to know from a term. In such a way, the coining process will become easier. Perhaps one day, this could be a way to register new concepts and new terms through the Internet.

Whether Termcoiner is successful or not depends on the well being of different knowledge bases and different processors. It may also depend on whether we are lenient in the number of alternative terms in the output. If we allow more, it is likely one or some of them are good. Unless there are significant statistics found, then we can have empirical ways to limit the number of alternative terms generated. It is never possible for any human terminologist to always arrive at a good new term at first sight, so we should not be so demanding of a computer system. The more interactive we can make Termcoiner to be at the first stage, the more likely it will not fail us too much.

Finally I like to conclude that the assumption that we start with English definitions need not always remain to be true. Good Chinese definitions could be provided instead as for our purpose. It is often the case that it is easier to translate a sentence than a term, and a paragraph than a slogan or motto, but we can start with the easier. However, it is still a matter of truth that we still rely on English to a great extent in the scientific world. Termcoiner is still at a preliminary design stage. It is an attempt to automate the coinage of new terms. With a proper attitude, we should pull the strengths of experts from different fields to assist in the research. The terms thus created can then be passed onto the users involved for acceptability test either using the top-down approach or the bottom-up approach, or both. Eventually, we will find the middle way to do it.

BIBLIOGRAPHY

Alshawi (1989) “Analyzing the Dictionary Definitions,” in Boguraev, B. & T. Briscoe (Eds.) Computational Lexicography for Natural Language Processing. London: Longman pp. 153-170.

Lu, Z. et al. (1977) Chinese Word Formation Principles, Hong Kong: Zhonghua Shuju.

Schulte, Rainer (1994) “Cross-cultural Communication on The Information Highway.” Translation Review. Texas: The University of Texas at Dallas.

Sinclair, John. (Ed.) (1987) Looking Up: An Account of the COBUILD Project in Lexical Computing, London: Collins ELT.

Sinclair, John. (1991) Word Formation. London: Harper Collins.

Sonneveld, Heimi B., & Kurt L. Loening (Eds.) (1993) Terminology: Applications in Interdisciplinary communication. Amsterdam: John Benjamins.

Yuan, C. et al. (1994) “The Construction and applications of Chinese Morpheme database,“. Manuscript.

APPENDIX A

A Terminology Builder - Termcoiner V1.0

System Architectura

****

LABURPENA / RESUMEN / RÉSUMÉ / ABSTRACT

Txinerarako terminologia-eraikitzaile automatizatua

Terminologia berria izan da edozein arlotako garapen azkarraren adierazlea, batez ere zientzien kasuan. Horren ondorioz, edozein diziplina bizik arazo ugari izaten ditu terminologia normalizatzean. Asmatzen diren terminoak sistematikotasunez sortu behar dira, esanahiari eta erabilerari dagokionean ezin dira anbiguoak izan, arlo bereko pareko beste termino batzuekin koherenteak izan behar dute. Hori guztia garrantzitsua da eta horrela egiten dela ziurtatu behar da. Hala ere, hizkuntza handi batek ez badauka berdintasunik (adibidez, kontinenteko Txinan, Taiwanen eta Hong Kongen erabiltzen den txinerak), diziplina batean erabiltzen diren terminoak termino-banku baten bidez bateratu behar dira. Termino-banku hori termino berriak erabiltzeko, biltzeko eta gordetzeko irizpideak ematen dituen erakunde batean egon beharko luke. Komunikazio honen gaia izango da zelan jarri martxan terminoen asmatzaile automatizatu bat eta termino-banku baten zati osagai bilakatu. Asmatzaile horrek datu-baseak erabiliko ditu, adibidez, dauden terminoen definizioen corpusak, arau sintaktikoen oinarria, lexikoiak (teknikoak eta orokorrak), morfemen datu-baseak (Yuan et al., eskuizkribua), txinerako hitz-eraketarako arauak (Lu 1975). Osagaiak hauek izango dira: egitura-aztertzaile bat, oinarritzat silabak hartzen dituen eraldatzaile bat (letraldaketa egiteko) eta asmatzaile bat. LDOCE esperientzian Alshawi-k erabilitako kontzeptuari jarraitzea eta hitzen ingelesezko definizioekin lan egitea da oinarrizko ideia. Definizio horien egitura-azterketatik gako-hitzak ateratzen dira, txinerara hitzez-hitzezko itzulpena egin ahal izateko. Gero, txinerazko gako-hitzen itzulpenak asmatzailean prozesatzen dira, txinerazko morfemen datu-baseak eta txinerazko hitz-eraketarako arauak erabiliz. Ondorioa itzulpen alternatiboak izango dira, gizaki adituek aukera dezaten.

Un constructor terminológico automatizado para el chino

La nueva terminología ha sido indicadora del rápido desarrollo, tanto en extensión como en profundidad, de cualquier campo que se encuentre en ebullición (sobre todo en las disciplinas científicas). Por consiguiente, toda disciplina que se encuentre de actualidad es bombardeada con problemas de estandarización terminológica. Es importante asegurarnos de que los términos acuñados son creados sistemáticamente, no son ambiguos ni en su significado ni en su uso y sí consecuentes con otros términos relacionados del mismo campo. De cualquier forma, en el caso de cualquiera de las lenguas más importantes en que la uniformidad no sea una norma (por ejemplo: el chino que se habla en China continental, en Taiwan y en Hong Kong), la unificación de los términos empleados en una disciplina depende de la existencia de un banco terminológico dentro de una organización donde el aportar una guía para el uso, recopilación y mantenimiento de nuevos términos debería ser uno de sus deberes cotidianos. Esta ponencia trata del cómo crear un acuñador terminológico automatizado y convertirlo en una parte integral del banco terminológico. El acuñador empleará bases de datos como por ejemplo corpora de definiciones de terminologías existentes, una base de reglas sintácticas, lexicones (técnicos y generales), una base de datos relativa a los morfemas (Yuan et al., manuscrito), reglas para la formación de palabras en chino (Lu 1975). Consistirá de un analizador gramático, un transformador basado en sílabas (con propósitos transliterativos) y un acuñador. Lo que

pretendemos es básicamente seguir la idea expresada por Alshawi en su experiencia LDOCE y trabajar con definiciones de términos en inglés. Las palabras clave se extraen del análisis de estas definiciones de manera que pueda llegarse a una traducción literal de las palabras clave del inglés al chino. Seguidamente, las traducciones de las palabras clave del chino son procesadas en el acuñador haciendo uso de la base de datos relativa a los morfemas en chino y de las reglas para la formación de palabras en chino. El resultado será una serie de traducciones alternativas para que personas expertas puedan escoger.

Un constructeur terminologique automatisé pour le chinois

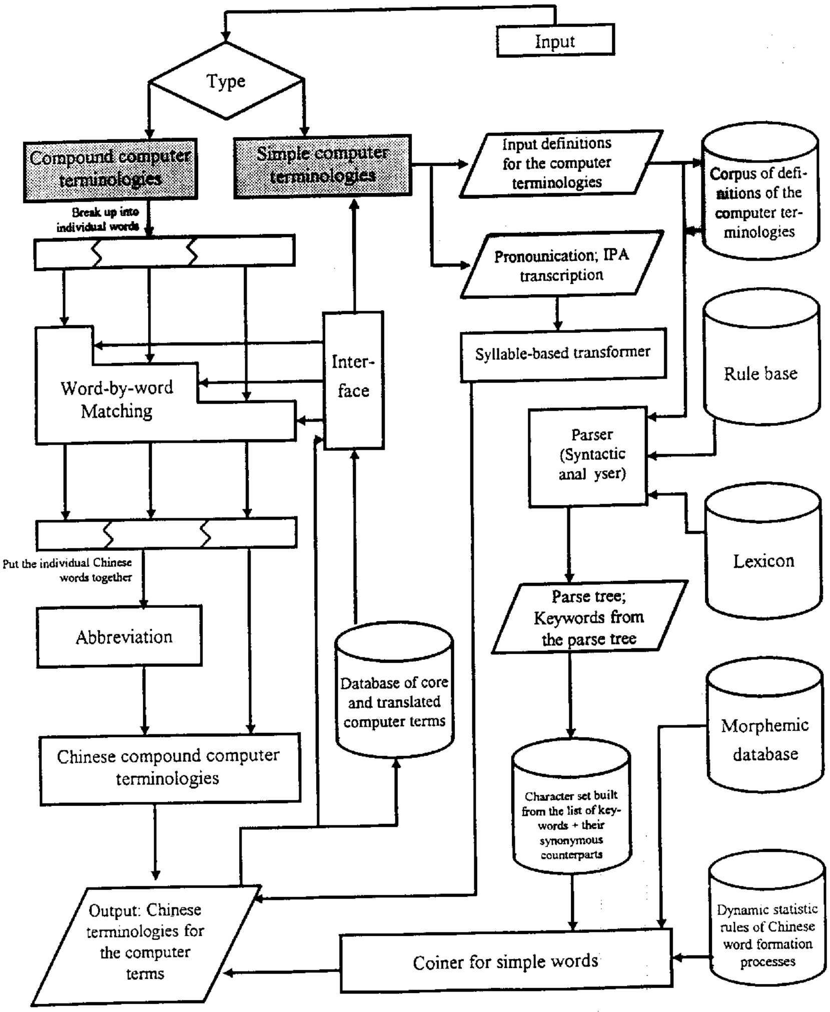
La nouvelle terminologie est un indicateur du développement rapide en long et en large de tout domaine d'actualité (notamment dans les disciplines scientifiques). Comme résultat, toute discipline active se voit bombardée par les problèmes de la normalisation de sa terminologie. Il est très important de s'assurer que les termes inventés sont créés systématiquement et qu'ils ne sont pas ambigus dans leur signification et dans leur utilisation, et qu'ils sont cohérents avec d'autres termes en relation avec le même domaine. Néanmoins, pour n'importe quelle langue majoritaire où l'uniformité n'est pas une norme (ex: le chinois tel qu'il est utilisé dans la Chine continentale, Taiwan et Hong Kong), le fait d'unifier les termes utilisés à l'intérieur d'une discipline dépend de l'existence d'une banque de termes dans une organisation, où l'orientation fournie dans l'usage, la collecte et l'entretien des termes nouveaux doit faire partie de ses obligations routinières. Le thème de l'exposé présent est: comment mettre en pratique un inventeur automatisé de terminologie, et de faire en sorte qu'il fasse partie intégrale d'une banque de termes. L'inventeur profitera des bases de données telles que les corps de définitions de terminologies existantes, la base de règle syntaxique, les lexiques techniques et généraux, les bases de données de morphèmes (Yuan et al., manuscrit), les normes de formation de mots chinois (Lu 1975). Il comprendra un analyseur grammatical, un transformateur basé sur les syllabes (pour des travaux de translittération) et d'un inventeur. L'idée de base est de continuer le travail d'Alshawi dans son expérience, ainsi que le travail de LDOCE en relation avec les définitions de termes en anglais. Les mots clés sont extraits de l'analyse grammaticale de ces définitions afin d'obtenir la traduction littérale de l'anglais au chinois. Par la suite, la traduction des mots clés en chinois sont traités dans l'inventeur en utilisant la base de données de morphèmes chinois et les normes de formation des mots chinois. Le résultat sera des traductions alternatives afin que les experts puissent faire leur choix.

Automated Chinese terminology builder

An increase in terminology has been an indicator of rapid development in breadth and depth in any hot field (especially in science and technology disciplines). As a result, any lively discipline is bombarded with problems in standardization of terminology. It is important to make sure the terms coined are systematically created, non-ambiguous in meaning and usage and consistent with other related terms in the same domain. However, for any major language where uniformity is not a norm (e.g. Chinese as used in mainland China, Taiwan, and Hong Kong), to unify the terms used within a discipline depends on the existence of a term bank in an organization where provision of guidance in usage and collection and maintenance of new terms should be its routine duties.

The theme of this paper is study how to facilitate automated coinage of terminology and make it an integral part of a term bank existing on the Internet. The terminology coiner will make use of databases such as corpora of definitions of existing terminology, syntactic rule base, lexicons (technical and general), morpheme database (Yuan et al., 1994), Chinese word formation rules (Lu et al. 1977). It will consist of a parser, a syllable-based transformer (for transliteration purposes) and a coiner. The basic idea is to follow Alshawi’s idea in their LDOCE experience and work with English definitions of terms.

Key words are extracted from parsing such definitions so that literal translation of English key words into Chinese can be achieved. Then the Chinese key word translations are processed in the coiner making use of Chinese morpheme database and Chinese word formation rules. The output will be alternative translations for human experts to choose from.


Input
Type
Compound computer terminologies
Simple computer terminologies
Break up into individual words
Word-by-word Matching
Inter- face
Put the individual Chinese words together
Abbreviation
Chinese compound computer terminologies
Database of core and translated computer terms
Output: Chinese terminologies for the computer terms
Input definitions for the computer terminologies
Pronounication; IPA transcription
Syllable-based transformer
Corpus of defi- nitions of the computer ter- minologies
Rule base
Parser (Syntactic anal yser)
Lexicon
Parse tree; Keywords from the parse tree
Morphemic database
Character set built from the list of key- words + their synonymous counterparts
Dynamic statistic rules of Chinese word formation processes
Coiner for simple words