Culling Scientific and Technical Terms from Text Corpora for Compiling a TermBank in Bangla

Niladri Sekhar Dash
Linguistic Research Unit Indian Statistical Institute Kolkata, India
Email: ns_dash@yahoo.com

ABSTRACT

In this paper I describe a few steps that we adopt to develop a digital TermBank after culling the Scientific and Technical Terms (STTs) from a text corpus of Bangla. Following the stages and methods of processing and analysis of corpus we are successful to develop a TermBank which now contains nearly 10,000 terms to be used in various works of linguistics and language technology. The strategy we use can be effectively applied on corpora of other Indian languages for same purposes. This confirms its utility and relevance in NLP works for Indian languages.

Keywords: *Scientific and technical terms, corpus, POS tagging, collocation, lemmatization, TreeBank, terminology, frequency*

1. Introduction

The development of a comprehensive digital database of scientific and technical terms (STTs) in a language is important in works of linguistics and language technology, such as, termbank compilation, linguistic resource generation, machine translation, machine learning, information retrieval, knowledge representation, text classification, language planning, online language education, dictionary compilation, text composition, and mass literacy (Sager 1994). Keeping such activities in mind, we have developed, as a project of our NLP activities, a comprehensive database of nearly 10,000 STTs extracted from a Bangla corpus of scientific texts compiled with data collected from the TDIL corpus developed for the language.

To be precise in presentation, we first define the concept of scientific term (Section 2) and technical term (Section 3) to draw a line of distinction between the two. Next, we describe methods we use to process the corpus (Section 4), and the architecture we use for TermBank compilation (Section 5). In conclusion (Section 6), we identify people who use this TermBank to address various needs of linguistics and language technology (Wright and Budin, 1997, pp. 370).

2. Scientific Term

The expression *scientific term* refers to single and multiword units that are used in different scientific texts in specialized senses. Although the literal meaning of the expression refers to specialized terms used in scientific texts, it is not confined to the fields of science only. Rather it encompasses all the specialized terms used in any discipline of human knowledge.

Based on the connotative meaning of the expression, *scientific terminology* refers to analysis of form, function, and usage of scientific terms in any area of human knowledge and information. Therefore, *scientific terminology*, in principle, investigates among other things, how various specialized terms have come to existence as well as their interrelationships within the fields of study. Thus, s*cientific terminology* refers to a more formal study, which systematically investigates *labelling* or *designating* of concepts specific to domains of human science. Moreover, it involves research and analysis of terms with a purpose of documenting and promoting their correct usage in various fields and domains. The study may be limited to one language or can more than one language at the same time to develop sharing portal for bilingual (or multilingual) scientific terminology database which has great relevance in a multilingual country like India where exchange and sharing of scientific knowledge and ideas are necessary for balanced progress of the country (Sonneveld and Loenning, 1994).

The importance of scientific terms is not limited to the area of information retrieval alone. It is also related to conveyance of concepts and meanings. Therefore, it is to be noted that the word *term* (i.e., *index terms*) that is used in the context of information retrieval is not the same as *term* that is used in the context of *scientific terminology*, since *term* in information retrieval does not always mean scientific terms of a discipline. In the realm of *scientific terminology*, the actual value of *scientific terms* is measured by their contexts of occurrence, classification, and their capability of explicating the senses they denote in a text. While investigating *scientific terms* we need to keep in mind the following issues (Resnik and Smith, 2003):

(a) Analysis of stock concepts used in a particular field of study,
(b) Identification of the terms assigned to the concepts,
(c) Compilation of scientific terms in the form of a database,
(d) Management of scientific term database for various access,
(e) Creation of new terms to accommodate new concepts,
(f) Establishment of conceptual correspondences among the scientific terms used in bilingual and multilingual texts.
(g) Sharing of scientific terms across languages to develop crosslingual TermBanks.

Whenever there arises a need for generating a database of *scientific terms* to address requirements of particular discipline, it becomes necessary to deal with several related issues, such as, plantation of new concepts, coinage of new terms (neologism), conceptualization of terms, identification of new meanings of existing terms, borrowing of terms, generation of acronyms, etc (Fernández, Colina and Peters, 2009). In fact, the conditions that are relevant for *neologism* are equally applicable to formation of scientific terms as both ask for careful consideration of several linguistic and functions, which are non-trivial at the time of coining new scientific terms. In this context, the observation of Barnhart (1978) appears to be quite relevant:

> "The vocabulary of science should be related to the general vocabulary of educated people so that the peculiar contributions of any science to our knowledge and understanding of the universe can be made a part of general knowledge. The basic terms of scientific and technical vocabulary should be so explained that the beginning student can comprehend them and relate them to his experience. It should be possible, both in

general purpose dictionaries and in specialized technical dictionaries to show that scientific terms are not merely hard words but results of a different and more exact structuring of the world by the scientist; parallel defining is of great importance as are cross references to closely related terms. The concept of *atom* is related to *molecule* and *nucleus* and *proton*; one term cannot really be understood without the others" (Barnhart, 1978, pp. 1927).

For many practical purposes, a bilingual (and multilingual) database of scientific terms is useful in manual and machine translation, teaching use of scientific terms in academic and translation schools, and in composition of scientific texts. At initial stage most of the scientific terms may appear quite incomprehensible to common users. It is, therefore, necessary to make these terms understandable to the common mass so that these become a part of the regular vocabulary of a language.

3. Technical Term

Scientific terms should not be confused with *technical terms* as these are different in denotation. Technical terms are used to define ideas and concepts within a special discipline or a field of speciality. *Technical terminology* is a specialized vocabulary or nomenclature of particular disciplines. These terms have specific definitions within the field, which is not necessarily the same as their meaning in common use. We can perhaps use the term *jargon*, which is nearest in sense; but the term *jargon* is more informal in definition and use, while *technical terms* have meanings strictly defined by the disciplines. So, by a simple definition, a *technical term* is a unique lexical entity, which has a truly specialized meaning within a specific field of human knowledge. It implies that a word or a phrase is highly typical within a particular field of study and only people directly linked to this field are familiar with these and use these. For instance, while *computer* is a scientific term, *firewall, toolbox, processor, CPU, RAM, software, hard drive, HTML,* etc. are technical terms understandable only to people related to it (Gomez and Gozalez, 2006, and Green, 1987).

A *technical term* evolves as a result of needs of the experts working in a discipline to communicate with utmost precision and brevity. It often has an effect of excluding the people who are not familiar with that particular specialized language of the discipline. It can cause severe difficulties for laymen. For instance a patient who fails to follow the discussions of the medical practitioners, cannot understand her own condition and treatment (Wanner, Bohnet, Giereth and Vidal, 2005).

A *technical term* should have the qualities to be scientifically accurate and intelligible. As we cannot do justice to definition of all technical terms in general, we can take help from experts of different fields for understanding these terms. It is better to have a board of experts of different fields who can advise us in the matter of defining technical terms. Since all technical terms do not find a place in a general dictionary, it is necessary to mark which technical terms are to be included in a general reference dictionary. However, in case of a *digital database of technical terms,* each and every technical term is included in the database. Newly coined technical terms as well as old technical terms are treated with equal importance (Justeson and Katz, 1995).

The issue of commonness and uncommonness of technical terms is addressed with consideration of goals of our general digital database. Even if some technical terms look artificial and ambiguous, these are included in our database and are given equal importance as the common ones. Moreover, in those cases where several terms denote one concept or one term denotes several concepts are also included in the database (Nash, 1993).

With regard to definition of technical terms, we argue that the definition of a technical term should be provided in such a manner that it is clearly intelligible to non-specialists as well. This will settle questions whether only a technical definition is to be given or a more general definition should also be there. For elucidation, consider the following examples:

Iron:

Technical Definition: It is the second heaviest stable isotope produced by the alpha process in stellar nucleosynthesis, made with a chemical element with symbol Fe (*ferrum*) and atomic number 26. It is a group 8 and period 4 element.

Formal Definition: It is a metal with lustrous and silvery color. It is the most abundant element in the core of meteorites and in the dense metal cores of planets such as Earth. It is one of the most common sources of ferromagnetic materials adopted in everyday use.

Coal:

Technical Definition: It is a fossil fuel formed in ecosystem where plant remains were preserved by water and mud through oxidization and biodegradation, and its chemical and physical properties have been changed as a result of geological action over time, thus sequestering atmospheric carbon. It is a readily combustible black or brownish-black rock, composed primarily of carbon and hydrogen along with small quantities of other elements, notably sulphur.

Formal Definition: It is a hard opaque black or blackish mineral or vegetable matter found in seams or strata below the surface of earth and used as a fuel and in manufacture of gas, tar, etc.

The examples given above show that definition of a technical term should be such formed that it is able to transmit ideas of science and technology in simple language for understanding of common mass. If it is not possible to provide precise definition of the terms, these should be explained with some equivalent illustrative terms, images, and pictorials so that the illustration becomes more expressive for understanding items or objects. By way of opting illustration (pictorial or otherwise) we are actually entering into a domain of encyclopaedic definition, which is necessary for most of the STTs of our database.

The development of a STT TermBank for Indian languages is still a far cry, even though, at least twenty years ago, digital text corpora for most of the Indian languages were made ready for NLP activities. Since there has hardly been any effort to develop TermBank in Indian languages, it is difficult to refer to any such work, although works are done in English and other languages (Daille, Gaussier and Lange, 1994; Ha, Fernandez, Mitkov and Corpas,

2008; Kim, Baldwin and Kan, 2009; Lefever, Macken and Hoste, 2009; Pantel and Lin, 2001). In the present context of NLP activities in Indian languages, we focus our attention towards this work for Bangla for the benefit of the language. The first task that we did in this case is to process the Bangla language corpus in various ways as proposed in the following sections.

4. Processing Language Corpora

Processing of corpus is mandatory for compiling a TermBank. For this we require some efficient techniques to process a corpus (manually or automatically) to fabricate a functional interface between input texts and resultant TermBank. The necessity for processing corpus arises after the accumulation of large amount of language data in electronic form based on which we devise techniques for processing texts for extracting relevant the terms along with contextual information (Macken, Lefever and Hoste, 2013). Although some corpus processing software are available for English, French, German, and other languages, for Bangla there is nothing. Therefore, keeping in mind the orthography (i.e., script) and nature of the Bangla language, we devise some processing techniques that are used for compiling the TermBank. The techniques that we use for processing the Bangla corpus include parts-of-speech tagging, concordance, collocation, lemmatisation, frequency sorting, and type-token analysis (Pala and Ganagashetty, 2012). Most of the techniques are also useful for other Indian language corpora for same purposes.

4.1 Part-Of-Speech (POS) Tagging

It is a type of text annotation, which involves attachment of special codes related to part-of speech of words used in corpus in order to indicate their lexico-syntactic features and functions they denote in texts. The part-of-speech assigned to the words is known as *tag*. When we tag words in corpus we follow a scheme for tagging part-of-speech to STTs in sentences. We do the work in 3 stages: pre-editing, tag assignment, and post-editing.

At pre-editing stage, we convert the corpus into a format suitable to assign part-of-speech to each word or a multiword combination. At the tag assignment stage, we assign only one POS tag to each STT with proper reference of its actual syntactic role in sentence. At the post-editing stage, we manually check all the POS tagged STTs to validate their semantic identity and tag assignment. We carry the work following BIS (Bureau of Indian Standard) tagset designed for the Indian language.

মহীভাবক আলোড়ন ভূ-পৃষ্ঠে লম্বভাবে (vertically বা radical direction) কাজ করে থাকে । এর ফলে ভূ-পৃষ্ঠের স্থানসমূহ খাড়াভাবে উপরে বা নীচে ওঠা-নামা করে । ব্যাপক আকারে মহাদেশ জুড়ে সাধারণত এই আলোড়ন বা আন্দোলন হয় বলে একে মহীভাবক আলোড়ন (Epeirogenic, গ্রীক শব্দ Epeiros অর্থে 'মহাদেশ' থেকে এসেছে) বলে । তবে মহীভাবক আলোড়নের প্রভাব স্থানীয়ভাবেই (locally) বেশি দেখতে পাওয়া যায় । এই আলোড়নের ফলে ভূ-পৃষ্ঠে প্রধানতঃ বিশালায়তন প্রাচীন মালভূমি (Shield) ও অন্যান্য মালভূমি এবং চ্যুতি বা স্রংসের ফলে গঠিত স্রংস ভৃগুতট (Fault Scrap), গ্রস্ত বা স্রংস উপত্যকা (Rift Valley) বা গ্রাবেন (Graben), স্তূপ পর্বত বা হোর্স্ট (Block Mountain বা Horst) ইত্যাদি নানাবিধ ভূমিরূপ গঠিত হতে দেখা যায় ।

Fig. 1: Raw text taken from the Bangla corpus

mahībhābak āloṛan bhū-pṛṣṭhe lambābhābe (vertically bā radical direction) kāj kare thāke. er phale bhū-pṛṣṭher sthānsamuha khāṛābhabe upare bā nīce oṭhā-nāmā kare. byāpak ākāre mahādeś juṛe sādhāranata ei āloṛan bā āndolan hay bale eke mahibhābak āloṛan (epeirogenic, grīk śabda epeiros arthe 'mahādeś' theke eseche) bale. tabe mahībhābak āloṛaner prabhāb sthānīyabhābei (locally) beśi dekhte pāoyā ýāy. ei āloṛaner phale bhū-pṛṣṭhe pradhānata biśālāyatan prāchin mālbhumi (shield) o anyānya mālbhumi ebaṃ cyuti bā sraṃser phale gaṭhita sraṃsa bhṛgutaṭ (fault scarp), grasta bā sraṃsa upatyakā (rift valley) bā grāben (graben), stup parbat bā horsṭ (block mountain bā horst) ityādi nanābidha bhūmirūp gaṭhita hate dekhā ýāy.

Fig. 2: Indic Roman version of the Bangla text

We discard multi-tier POS annotation scheme to restrict us to single-tier tag assignment as the present work does not require multi-layered POS information for nouns used in the corpus. Moreover, only those nouns that are eligible to be considered as STTs are tagged in the text (Ekbal and Bandyapadhyay, 2008). Given below is a sample text of Bangla corpus in original Bangla script (Fig. 1), in Indic Roman script (Fig. 2) and its POS tagged version (Fig. 3). Nouns are POS tagged following the BIS tagset stated above.

mahībhābak/NN/ āloṛan/NN/ bhū-pṛṣṭhe/NN/ lambābhābe (vertically bā radical direction) kāj kare thāke. er phale bhū-pṛṣṭher/NN/ sthānsamuha/NN/ khāṛābhabe upare bā nīce oṭhā-nāmā kare. byāpak ākāre mahādeś/NN/ juṛe sādhāranata ei āloṛan/NN/ bā āndolan/NN/ hay bale eke mahibhābak/NN/ āloṛan/NN/ (epeirogenic, grīk/NN/ śabda epeiros/NN/ arthe 'mahādeś'/NN/ theke eseche) bale. tabe mahībhābak/NN/ āloṛaner/NN/ prabhāb sthānīyabhābei (locally) beśi dekhte pāoyā ýāy. ei āloṛaner/NN/ phale bhū-pṛṣṭhe/NN/ pradhānata biśālāyatan prāchin mālbhumi/NN/ (shield/NN/) o anyānya mālbhumi/NN/ ebaṃ cyuti/NN/ bā sraṃser/NN/ phale gaṭhita sraṃsa/NN/ bhṛgutaṭ/NN/ (fault/NN/ scarp/NN/), grasta/NN/ bā sraṃsa/NN/ upatyakā/NN/ (rift/NN/ valley/NN/) bā grāben/NN/ (graben), stup/NN/ parbat/NN/ bā horsṭ/NN/ (block/NN/ mountain/NN/ bā horst/NN/) ityādi nanābidha bhūmi-rūp/NN/ gaṭhita hate dekhā ýāy.

Fig. 3: POS tagged version of the Bangla text

When we look into the POS tagged version we find that there are some compound STTs the formative constituents of which are written separately with a space or a hyphen in between, as the following examples show (Fig. 4):

bhū-pṛṣṭhe/NN/ (earth surface)
bhū-pṛṣṭher/NN/ (of earth surface)
mahibhābak/NN/ āloṛan/NN/ (tectonic movement)
cyuti/NN/ sraṃser/NN/ (of fault scrap)
sraṃsa/NN/ bhṛgutaṭ/NN/ (scarp rift)
grasta/NN/ upatyakā/NN/ (rift valley

sraṃsa/NN/ upatyakā/NN/ (scrap valley)
stup/NN/ parbat/NN/ (block mountain)

Fig. 4: Detached compound STTs in Bangla corpus

Such compound STTs demand additional attention at the time of termbank compilation because while the first term (i.e., T_1) remains free from inflection, second term (i.e., T_2) carries inflection and both the terms constitute together to form a single SST, e.g.,*hū-pṛṣṭher*/NN/ (of earth surface) *cyuti*/NN/ *sraṃser*/NN/ (of fault scrap), etc. We adopt a strategy of chunking the constituent members together within square brackets to couple and put such detached STTs together, as shown below (Fig. 5).

[bhū-pṛṣṭhe]/NN_CMP/
[bhū-pṛṣṭher]/NN_CMP/
[mahibhābak āloṛan]/NN_CMP/
[cyuti sraṃser]/NN_CMP/
[sraṃsa bhṛgutaṭ]/NN_CMP/
[grasta upatyakā]/NN_CMP/
[sraṃsa upatyakā]/NN_CMP/
[stup parbat]/NN_CMP/

Fig. 5. Chunked compound STTs in Bangla corpus

To deal with inflected STTs, we first thought to run process of lemmatization on the corpus (discussed at section 4.4) before POS tagging is carried out. But we give up this idea, because if we do so, then all detached multiword STTs will be decomposed into two separate terms due to which T_1 and T_2 will be listed as two separate terms. This is a distortion of actual language data as well as non-reliable representation of lexical information of a language. This implies that unless detached multiword STTs are chunked before as single word units, there is a chance of having serious mistakes in lemmatization.

When we look at the POS-tagged sample given above, we find some important information about the type or genre of the text as presented below:

(a) Because of the presence of a large number of STTs of a particular subject area in the corpus, it is easy to identity to which discipline the text belongs (i.e., geology).
(b) Each STTs in the text is identified and tagged with a specific POS tag.
(c) The STTs are tagged in accordance to their grammatical role in sentence. If it is not done this way, we fail to understand actual syntactic-cum-semantic role of the STTs in the text.
(d) If STTs are not previously tagged at grammatical level, it is not possible to identify their correct POS tag as well as their contextual meanings— which can have reverse effect at the time of lemmatization.
(e) In the POS tagged corpus there are examples of compound STTs where constituents are written as separate terms (e.g., *[stup/NN/ parbat/NN/]/NN_CMP/]*. If such a STT is not identified and tagged properly as a single word unit, there is a chance that it will be decomposed into two separate terms, and as a result of this, the actual lexical

information of the compound STT will be lost and lemmatization process will yield wrong outputs. These issues guide us implement POS tagging first on the corpus before the words are put to lemmatization.

4.2 Concordance

Concordance is a process of indexing words used in a text. In concordance words are indexed with close reference to the place of their occurrence in text to show their possible range of usage varieties in text. Also it helps to understand the distributional and semantic patterns of words and terms collected from corpora in a desired fashion for subsequent analysis and observation. Introduction of computer has made concordance an easy process to compile and arrange words in a desired manner. In case of a TermBank, it is indispensable in understanding actual contextual identity of STTs, because it gives a scope to access the possible patterns of use of STTs in a text. Since it displays total number of use of a STT—each one in its own contextual environment— it becomes easy to understand in which sense a STT is actually used in a text. Due to flexibility in operation, determination of contextual frame of a STT may, however, vary depending on various criteria, such as, fixed number words on either side of a target SST, finding the sentence boundaries of SST, etc. (Chang, 2005).

The application of concordance on corpus yields varieties of information, which are largely responsible for tracking the sense denotation (if any) of STTs based on the type of a text (subject area). Such information is not available via intuition or in a dictionary. Due to excellent advantage, concordance is often used on corpus to search out single and multiword units as well as STTs along with contexts of their occurrences. With the help of concordance, it is not difficult to examine all the varieties of occurrence of different STTs in texts. For instance, in the table below (Table 1), we cite a sample concordance list of the term *software* to show how this new technical term is used in *British National Corpus* and how it varies in sense due to its occurrence in different contexts.

The use of concordance on Bangla corpus helps us identify STTs in their complete syntagmatic and paradigmatic details. With options open for left and right hand sorting, it is quite useful for us to investigate if any STT is polysemous in nature with a wide range of senses. In fact, it gives us a scope to access STTs in multiple syntactic frames to construct their multilayered semantic and functional profiles. In essence, concordance offers a unique opportunity to test, analyse, and document similarities and differences of STTs in the texts (Wright and Budin, 1994).

the application of	software	such as word processor
Firmware is a	software	that is programmed to
methods to test that a	software	is a fit product before it is
utilities and application of	software	that serve in combination
for direct application	software	or subsets thereof we need
in computer technology	software	are often regarded as one
these types of	software	include web pages and all
at the lowest level	software	consists of machine reading
in computer science	software	engineering software is all

basis for most modern	software	was first proposed by
on generally used	software	systems on the desktop
computer system divide	software	system into three classes
the purpose of systems	software	is to unburden applications
the programming	software	usually provides tools and
there are three layers of	software	performing variety of tasks
usually a platform	software	often comes bundled with
to change the platform	software	the operational modalities

Table 1: Concordance of *software* taken from an English text

4.3 Collocation

Collocation is a well-known linguistic phenomenon which is discussed with evidences carefully selected from many language texts. It is defined as occurrence of two or more words within a short space of each other in a text (Sinclair, 1991, pp. 170). The technique for identifying collocation in a text is important for evaluating relevance of consecutive occurrence of any two words (or STTs) in a piece of text. In return, it projects into the functional nature of the lexical items used in a text and reflects on "interlocking patterns of the lexis" in text (Williams 1998). While studying the corpus we are interested to know to what extent the actual patterns of use of STTs differ from the patterns that have been expected to form (Barnbrook, 1998). This query is linked to our assumption that mental lexicon is made up not only with single word units but also with multi-word units— both fixed and variable.

When we run collocation on Bangla corpus, it yields various kinds of information about the nature of collocation of STTs in the text. In fact, a systematic analysis of collocations of STTs in the corpus help us indentify some STTs that often take part in collocation as well as to understand how they perform based on their collocational usage. Moreover, it supplies vital information about their patterns of lexical association, which we require to analyse to understand their nature of sense denotation through collocation.

What we understand from our study on collocation of STTs in the Bangla corpus is that a list of collocation not only includes information about the frequency of use of STTs in collocation but also generates specific statistical counts to prepare a frame on their collocation patterns. In fact, without reference to their frequency of collocation in texts it is difficult to understand the finer aspects relating to their distribution, their possible sense variations, and distinction among the senses they denote while these are positioned in different collocation frames.

The analysis of collocation also shows us that by referring to contexts we empirically determine which pair of STTs maintains substantial collocation relationship between them. The most suitable formula we use is *Mutual Information* (MI) which helps to compare the probability of any two STTs (STT_1 & STT_2) occurring together as an attested lexical event against their probability of use as a result of chance. For each pair of STTs, we take some statistical scores from the corpus to conclude that where there is a higher score, there is

greater possibility of their collocation. Thus, reference to MI becomes useful in evaluation of occurrence of collocation of STTs used in a language.

The hand-on experience of collocation analysis of STTs used in corpus leads us realise the functional relevance of collocation is the works of STT database compilation, TermBank generation, development of lexical profile of STTs, etc. It also helps us understand the followings:

(a) MI of collocation helps to find mulltiword STTs from corpus to compile separate STT databases to build a database of translational equivalents, analyse patterns of collocation of STTs, and design materials for language education.
(b) It helps to group all multiword STTs in separate database to identify the range of their sense variation and to know how they generate new senses by collocating with new terms.
(c) Identification and analysis of collocation of STTs helps to understand and identify their differences in usage as well as their collocation-based sense denotation.
(d) It helps to understand the nature and pattern of semantic linkage between two synonymous STTs.
(e) Patterns of collocation of STTs in corpus show that they possess notable differences in lexical association resulted from differences in distribution across discourse types.

In essence, analysis of examples of collocation of STTs in text shows that they are rarely equivalent in sense and function when considered in terms of their distribution in texts. Thus, MI regarding delicate differences of collocation of STTs become important for us in terminology database generation, machine aided translation, language processing, dictionary compilation, and language teaching.

4.4 Lemmatisation

In linguistic analysis the term *lemma* refers to the basic form of a word disregarding its grammatical change, i.e., tense and plurality (Biber, Conrad and Reppen, 1998, pp, 29). In language processing, lemmatisation extracts a base from an inflected (or affixed) word— the headword that we look for in a dictionary. For tasks such as frequency counts, TermBank compilation, base form generation, etc., lemmatization is an indispensable technique by which we can group together different forms of an inflected word so that we can collectively display them under one head (Barnbrook, 1998, pp. 9). In works of vocabulary study, lexicology and dictionary compilation lemmatization allows us produce frequency and distribution information for inflected words as well as lemmas (Sánchez and Gomez, 1997). In the table below (Table 2) we present a list of inflected STTs that are collected and compiled from the text given above (Fig. 4) and their lemmatized forms to give an idea how inflected STTs are converted into lemmas.

Inflected STTs:

bhū-pṛṣṭhe/NN/ bhū-pṛṣṭher/NN/ sthānsamuha/NN/ ākāre/NN/ arthe/NN/ āloṛaner/NN/ bhū-pṛṣṭher/NN/ mālbhumir/NN/ mālbhumite/NN/ cyutir/NN/ sraṃser/NN/ bhṛgutaṛer/NN/ upatyakāy/NN/ grābene/NN/ stupe/NN/ parbater/NN/

SSTs	POS	Suffix	Lemma
bhū-pṛṣṭhe	NN	-e	bhū-pṛṣṭh(å)
bhū-pṛṣṭher	NN	-er	bhū-pṛṣṭh(å)
sthān samuha	NN	-samuha	sthān
ākāre	NN	-e	ākār
arthe	NN	-e	arth(å)
āloṛaner	NN	-er	āloṛan
bhū-pṛṣṭher	NN	-er	bhū-pṛṣṭh(å)
mālbhumir	NN	-r	mālbhumi
mālbhumite	NN	-te	mālbhumi
cyutir	NN	-r	cyuti
sraṃser	NN	-er	sraṃs(å)
bhṛgutaṭer	NN	-er	bhṛgutaṭ
upatyakāy	NN	-(ā)y	upatyakā
grābene	NN	-e	grāben
stupe	NN	-e	stup
parbater	NN	-er	parbat

Table 2: Lemmatization of words in the Bangla corpus

In principle, lemmatization allows us extract and examine all the variants of particular LEMMAS of the inflected words to produce frequency and distribution information for them. At the time of STT termbank compilation it helps us count the possible numbers of inflected forms generated from specific STTs. It also helps us identify those STTs, which are inflected in form as well as know how many times these are inflected and in which ways. The examples listed above (Table 2) show how a few STTs (e.g., *mālbhumi*, *bhū-pṛṣṭha*, etc.) are inflected with different inflections in texts and how lemmatization process removes the inflection part from the STTs to generate a lemma list.

4.5 Frequency Sorting

The frequency of use of STTs in a corpus is of great value in collection of STTs terms as well as in compilation of a TermBank as it is necessary to know which STTs register high frequency while others are sparse in use. In case of Bangla corpus all the STTs are arranged in accordance with their frequency of use to identify which STTs are most frequent and which are least in use in the language. The list is arranged in ascending and descending order based on requirements of terminologists, who are going to use the TermBank to compile a dictionary of STTs.

Generally, as we understand, a small-sized corpus provides too small a list of STTs to be interesting and useful. But a corpus of millions of words produces a frequency list that is useful for gathering detailed terminological information of a particular text, since listed STTs become comparable to a large population belonging to same genre or type of text for statistical validation. Moreover, as most of the frequently used STTs show variety in distribution in texts, many marked changes in the patterns of their distribution become significant in linguistic analysis and generalisation. For instance, we note that while highly

frequent STTs are easily found in a small corpus, less frequent STTs do not occur unless the corpus is big enough with huge amount of text samples obtained from different sources.

Since a frequency list can provide vital cues to know which STTs occur in different frequencies in a text, by examining a frequency list, we can collect the most frequently used STTs to compile *graded terminology databases* that can be handy for teaching technical terms to learners. Moreover, information of frequency of use of STTs is useful in dictionary compilation, terminology database generation, and term-based text materials composition. In the task of developing dictionaries of STTs, frequency information can lead us to decide which STTs will be included and which STTs will be ignored, as the most frequent STTs get priority over the rarely used ones. Thus, frequency of use of STTs helps us enlist and accept the most frequent STTs so that we can elaborate these terms to make them popular among the language users.

Further application of frequency information of STTs may be visualised in measuring the range of sense variation of the STTs. Since the most common senses of the STTs occur more often than their least common senses, it is necessary to identify the senses of STTs and select the most frequent common senses for reference and use in dictionaries and language teaching resources (Wills, 1990, pp. 122). In essence, frequency information of senses of STTs can establish their functional relevance in STT analysis, description, sense disambiguation, and usage documentation.

Since most of the existing Indian language corpora are not yet used for domain-based compilation of STTs, we think, it is necessary to focus our attention towards this direction so that we can develop TermBanks for most of the Indian languages to be used in machine (and manual) translation and other language processing works. However, we must keep in mind the problems relating to proper identification of STTs used in Indian language corpora. Otherwise, we shall make false observations and wrong deductions about the STTs and their linguistic information.

4.6 Type-Token Analysis

The last indispensable stage of TermBank compilation from a corpus is type-token analysis of STTs found in the corpus. It is made up with following two basic steps: (a) alphabetical sorting of STTs, and (b) removal of multiple tokens after storing a type of a STT. Once the STTs are obtained from the corpus, they are passed through the stage of alphabetical sorting so that the STTs are arranged in alphabetical order. Alphabetical sorting is a list of terms, which are arranged in alphabetical order with a tag denoting their frequency of use in the corpus. Since it is used for general reference purpose, it plays secondary role in the context of checking frequency of particular terms in the text. However, it is useful as an object of independent study because it helps in formulation of hypotheses and for checking the assumptions that have been made before (Kjellmer, 1984). In alphabetical sorting, STTs are displayed in a vertical form for general reference purpose and the list is formed in such a way that each STT is produced in a separate line for better representation and comprehension.

The next stage is related to removal of multiple tokens from the list of STTs. Since an alphabetically sorted list may contain multiple entries of same STT (i.e., *tokens*), we remove

identical terms from the list after storing one of the variants (i.e., *types*) as a representative of the tokens. For instance, an alphabetically sorted list may contain ±100 tokens of a single term (e.g., *isotope*)— in inflected or non-inflected form. Our primary task is, therefore, to preserve only one of the tokens as a 'type' and remove other forms from the list. Other forms are removed as these are identical replicas of the type selected for the list. By this process, a large list of tokens is reduced to a small and manageable set of types, which we finally preserve in the final database of STTs for the language.

5. The STTS TermBank

While compiling STTs TermBank for Bangla, we distinguish two types of TermBank, based on the nature of accumulation (Condamines, 2010):

(a) **Selective TermBank:** This contains only a few or a limited number of STTs with a goal for using them in translation works by professionals. The translators, while translating the scientific and technical texts, often seek for specific STTs or a group of terms to solve particular translation problems.

(b) **Exhaustive TermBank**: This contains all the STTs that we are able to collect from corpus. It is large in size capturing almost all the terms of a specific area of human knowledge. It is exhaustive as far as availability of STTs in the corpus is counted. Although exhaustive it is not complete as coining new STTs is a continuous on-going process of a language.

Theoretically, it is possible to postulate three methods with regard to the manners of accumulation of STTs from a corpus:

(a) **Selection Method:** In this method we collect STTs from books and journals of different domains and disciplines, covering almost all major domains of human knowledge. We identify and select relevant STTs manually to compile a database. In this case, however, we have to go through different types of text to cull STTs we think appropriate to be included in the database.

(b) **Collection Method:** In this method we collect STTs terms from earlier works already available in a language, such as dictionaries, word books, books of terms, and word books of specific fields and disciplines. Here, although our task is not exhaustive, it asks for careful collection of the terms suitable for the database.

(c) **Compilation Method:** This is a robust and useful method of TermBank development. Here we directly use text corpus of different disciplines and subject areas, process it in different ways (discussed above), and extract STTs with all necessary information required to be furnished in a TermBank.

The primary activities that we carry out to extract STTs from the Bangla corpus are depicted below (Fig. 6).

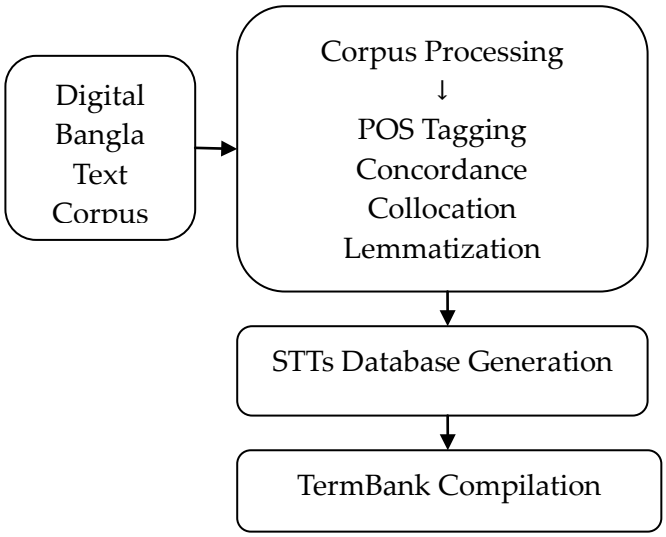

Fig. 6: Compilation of STTs TermBank from corpus

6. Conclusion: STT TermBank Users

It is necessary to identify users of STT TermBank, which we are developing through several stages of corpus generation and processing. In our opinion, the STT TermBank developed in this fashion, may be used by four types of users, namely, language specialists, content specialists, language technologists, and general people (Temmerman, 2000, pp. 26).

Among language specialists, lexicographers and thesaurus compilers require detailed information of general and specific use of STTs to develop lexical databases, term dictionaries, thesauri and reference materials. Terminologists and technical writers require this database and supporting information to standardise technical terminology as well as to increase general stock of terminology of a language. They access these terms to investigate linguistic phenomena of diverse kinds and to verify evidence of their own or their informants. Language teachers can use this database in terminology teaching, technical course book preparation, grammar writing, and similar other works.

Among content specialists, historians need this database to study a language through elaborate analysis of terms used in texts. They also use this database to discover implicit marks of time in obscured documents. Literary critics may use these terms in their research into stylometrics and stylostatistics as statistical information about use of STTs plays crucial role in determining ascription of dubious works to unknown authors. They also use STTs stamped with statistical information to identify different types of text based on density of use of STTs in texts.

Among language technologists, information retrievers may use TermBank to devise mechanisms for extracting information from large body of texts to build lexical knowledgebase, find information of terms for indexing, and to summarise important content of texts. Also they use STTs database to test presence or absence of regularities of use of terms in texts. Translators may utilize STTs TermBank as a necessary inputs to develop

bilingual and multilingual translational equivalents required in manual and machine translation.

Finally, general language users may like to refer to TermBank for language description, language study, language cognition, text composition, language therapy, and publication.

References

Barnbrook, G. (1998) *Language and Computers.* Edinburgh: Edinburgh University Press. P. 87.

Barnhart, C. (1978) American lexicography 1945-73. *American Speech*. 53(2): 83-140.

Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics - Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Chang, J.S. (2005): Domain specific word extraction from hierarchical web documents: a first step towards building lexicon trees from web corpora. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.* Pp. 64-71.

Condamines, A. (2010) Variations in terminology: application to the management of risks related to language use in the workplace. *Terminology.* 16(1): 30–50.

Daille, B., Gaussier, E. and Lange, J-M. (1994) Towards automatic extraction of monolingual and bilingual terminology. In: *COLING 94, 15th International Conference on Computational Linguistics, Proceedings.* Kyoto, Japan, pp. 515-521.

Ekbal, A. & Bandyapadhyay, S. (2008) Web based Bangla news corpus for lexicon development and POS tagging. *POLIBITS*, 37(1): 20-29.

Fernández, T., Colina, M.A.F. & Peters, P. (2009) Terminology and terminography for architecture and building construction. *Terminology.* 15(1): 10-36.

Gomez, A. and Gozalez, J. (2006) Meaning and anisomorphism in modern lexicography. *Terminology.* 12(2): 215-234.

Green, J. (1987) *Dictionary of Jargon.* London: Routledge & Kegan Paul.

Ha, L., Fernandez, G., Mitkov, R. & Corpas, G. (2008) Mutual bilingual terminology extraction. In *LREC 2008: 6th Language Resources and Evaluation Conference.* Marrakech, Morocco, pp. 1818-1824.

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

Kim, S., Baldwin, T. & Kan, M-Y. (2009) An Unsupervised Approach to Domain-Specific Term Extraction. *Proceedings of the Australasian Language Technology Association Workshop* 2009. Sydney, Australia, pp. 94- 98.

Kjellmer, G. (1984) Why ‘great: greatly’ but not ‘big: bigly? *Studia Linguistica.* 38(1): 1-19.

Lefever, E., Macken, L. & Hoste, V. (2009) Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus. In EACL’09 *Proceedings of the 12th Conference of the European Chapter of Association for Computational Linguistics.* Athens, Greece, pp. 496-504.

Macken, L., Lefever, E. & Hoste, V. (2013) Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology.* 19(1): 1-30.

Nash, W. (1993) *Jargon: Its Uses and Abuses.* Oxford, UK: Blackwell.

Pala, K. & Ganagashetty, S.V. (2012) Challenges and opportunities in automatically building bilingual lexicon from web corpus. *Interdisciplinary Journal of Linguistics.* 5(1-2): 169-184.

Pantel, P. & Lin, D. (2001) A Statistical Corpus-Based Term Extractor. In: Stroulia, E. and S. Matwin (eds.) *Advances in Artificial Intelligence, 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence,* AI 2001, Ottawa, Canada, Proceedings. LNCS vol. 2056. Berlin: Springer, pp. 36–46.

Resnik, P. & Smith, N.A. (2003) The web as a parallel corpus. *Computational Linguistics.* 29(3): 349-380.

Sager, J.C. (1994): Terminology: custodian of knowledge and means of knowledge transfer. *Terminology.* 1(1): 7-15.

Sánchez, J.A. & Gomez, P.C. (1997) Predictability of word forms (types) and lemmas in linguistic corpora: a case study based analysis of the CUMBRE corpus: an 8-million-word corpus of contemporary Spanish. *International Journal of Corpus Linguistics.* 2(2): 259-280.

Sinclair, J.M. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sonneveld, H. & Loenning, K. (1994) Introducing terminology. *Terminology.* 1(1): 1-6.

Temmerman, R. (2000) *Towards new ways of terminology description: the sociocognitive-approach.* London: John Benjamins.

Wanner, L., Bohnet, B., Giereth, M. & Vidal, V. (2005) The first steps towards the automatic compilation of specialized collocation dictionaries. *Terminology.* 11(1): 143-180.

Williams, G.C. (1998) Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics.* 3(1): 151-172.

Wills, J.D. (1990) *The Lexical Syllabus.* London: Collins.

Wright, S.E. & Budin, G. (1994) Data elements in terminological entries: an empirical study. *Terminology.* 1(1): 41-59.

Wright, S.E. & Budin, G. (1997) *Handbook of Terminology Management, Vol. 1, Basic Aspects of Terminology Management*, Amsterdam: John Benjamins.