Designing a Generic Scheme for Etymological Annotation (EA): A New Type of Language Corpora Annotation

Niladri Sekahr Dash
Linguistic Research Unit
Indian Statistical Institute, Kolkata
ns_dash@yahoo.com

Mazhar Hussain
Centre for Indian Languages
Jawaharlal Nehru University, New Delhi
mazharmehdi@gmail.com

Abstract

We have introduced here a new type of corpus annotation which we call **Etymological Annotation (EA)**. We propose this new type because although, over the years, scientists have proposed corpus annotation of various types (Atkins, Clear and Ostler 1992, Biber 1993, Leech 2005), nobody has ever suggested that words included within corpora need to be annotated at their etymological level so that one can retrieve necessary linguistic information relating to antiquity of words and terms used in corpora. The applicational relevance of etymologically annotated corpora may be visualized in language description, language planning, language education, lexicology, language technology as well as in compilation of general, historical, learner and special dictionaries. In case of those languages, where one comes across large number of words borrowed from neighbouring and foreign languages, the proper identification of source of origin of words carries tremendous referential relevance in cross-lingual lexical database generation, morphological processing, part-of-speech tagging, e-learning, digital lexical profile generation, information retrieval, machine learning, and language documentation. Thus, etymologically annotated corpora become an essential resource of applied linguistics and language technology. We propose here to define this new event with necessary direction and guidance to develop etymologically tagged language corpora for all natural languages.

1 Introduction

A simple look at the vocabulary of any natural language will invariably show that a large part of its vocabulary is actually obtained from foreign languages, besides having its own lexical stock inherited from native ancestral languages. Also analysis of the lexical stock will show that most of the words are naturalized to such an extent that it is almost difficult to trace their source of origin (Dash, Dutta Chowdhury and Sarkar 2009). This leads us to introduce the concept of **Etymological Annotation (EA)** where the basic task is to tag etymological information to each and every word and term used within corpora with regard to its source of origin (or antiquity) for future reference and application.

In our assumption, EA on corpora, in the long run, will become simply indispensable for each natural language, because the event of lexical borrowing is an inevitable linguistic phenomenon through which each natural language passes through for its continuous growth and survival. In fact, many advanced languages like *English, German, Spanish, French, Italian, Japanese, Chinese, Portuguese, Hindi, Bengali, Marathi, Tamil, Telugu,* etc. which are proud of having large pool of vocabulary, gladly admit the truth that much of their vocabulary are obtained from other languages – both native and foreign. For instance, Hindi language has a large stock of words in its vocabulary and a major part of it is obtained from *Sanskrit, English, Arabic, Persian, Spanish, German, Urdu, Punjabi, Gujarati, Kashmiri, Bengali,* etc. However, there is hardly any well-documented record (for most of the languages) to show which lexical items are inherited or borrowed from which languages into the vocabulary of a language, it is difficult for investigators to find how the vocabulary of a language has evolved across space and time on different diachronic scales.

Here arises the functional relevance of EA on language corpora (Leech and Fligelstone 1992). The annotation scheme proposed in this paper can solve the problem of etymological indeterminacy with proper documentation of etymological information for each and every word used in a piece of text, as each and every word in the corpora is annotated with a specific tag of its source language. The process may be initially carried out manually for developing a trial database for machine learning as well as tagging automatization in subsequent stages. The ultimate goal is to develop a system for automatic EA of text corpora of a language utilizing information and knowledge found from a supervised machine learning system.

Keeping several issues of EA in mind, we have briefly referred to various types of annotation (Section 2); noted the state-of-the-art of annotation in Indian languages corpora (Section 3); made attempt for etymology-based vocabulary classification (Section 4); defined an elaborate tagset for EA (Section 5); discussed the methods we have adopted for EA (Section 6); and finally reported some findings obtained from an etymologically annotated corpus (Section 7). The applicational importance of EA corpora is elaborated in conclusion.

2 Types of Corpus Annotation

In a broad sense language corpora can have two types annotation: (a) intralinguistic annotation, & (a) extralinguistic annotation (Dash 2011). While *intralinguistic annotation* involves encoding words, terms, phras-

International Joint Conference on Natural Language Processing, pages 64–71,
Nagoya, Japan, 14-18 October 2013.

es, and other linguistic items used within corpora with their part-of-speech and/or morpho-grammatical information; *extralinguistic annotation* encodes same linguistic items with information relating to their orthography, meanings, discourse, pragmatics, anaphora, and sociolinguistics (Leech and Wilson 1999, Sperberg-McQueen and Burnard 1994, Smith, Hoffmann and Rayson 2007). Thus, based on the nature of information tagged with words and terms used within corpora, annotation are classified into 6 major types, namely, *orthographic annotation, prosodic annotation, grammatical annotation, semantic annotation, discourse annotation,* and *anaphoric annotation*.

(a) **Orthographic Annotation**: It represents a text, as much as possible, as it actually exists in its complete natural state, despite attachment of multiple extratextual and intratextual tags (Dash 2011). It tags, for example, different orthographic symbols, such as, *single quotes, double quotes, type size, indentation, bold face, italics,* etc. as well as *capital letters, pauses, periods, apostrophes, segments, paragraphs, lines, punctuations, abbreviations, postcodes,* etc. used in a piece of text (Sperberg-McQueen and Burnard 1994).

(b) **Prosodic Annotation**: It is carried out on a spoken text corpus after a speech corpus is transcribed into its written form (Johansson 1995). In general, it tags all kinds of prosodic features, such as, *pitch, loudness, length, pause, tone, intonation variation, accent, juncture,* and other suprasegmental features and properties observed in spoken text (Grice, Leech, Weisser and Wilson 2000).

(c) **Grammatical Annotation**: It involves assigning specific part-of-speech to words after understanding their actual grammatical roles within a given text (Greene and Rubin 1971). At sentence level, this information may be tagged for chunks such as multiword expressions, local word groups, phrases, and idiomatic expressions, etc. (Francis 1980, Garside 1987, DeRose 1988). It may also involve marking of dependencies, constituents, named entities, and predicates and their arguments found within sentences (Kupiec 1992, Smith and McEnery 2000).

(d) **Semantic Annotation**: It is used on corpora to tag appropriate sense a particular word denotes within a given context (Löfberg, Juntunen, Nykanen, Varantola, Rayson and Archer 2004). The basic goal is to distinguish primary lexicographic senses of words – a process used in word sense disambiguation and assignment of semantic domains to words used in texts (Löfberg, Archer, Piao, Rayson, McEnery, Varantola and Juntunen 2003). It tries to identify the semantic information of words as well as exhibits semantic relationships underlying between words within texts. It also tags agent-patient relationships of words denoting their particular actions (Löfberg *et al.* 2005, Piao *at al.* 2005, Piao *at al.* 2006).

(e) Discourse Annotation: It tags discourse elements, sociolinguistic cues, pragmatic features, and other extralinguistic features found embedded within a piece of text (Archer and Culpeper 2003). Corpora are annotated beyond sentence boundaries to explore discourse as well as pragmatic relations expressed by linguistic elements used in corpora (O' Donnell 1999). It is argued that proper identification of discourse elements in spoken texts is indispensable for indicating conversational structure of dialogic interaction in case of normal speech events (Stenström 1984).

(f) **Anaphoric Annotation**: It tries to identify different types of anaphora used in texts as well as lists and sorts these forms to dissolve anaphoric complexities. It tags anaphora and anaphoric relations of words used within a text for intra-sentential or intra-textual references. Usually, various pronouns and nouns are co-indexed within a broad framework of cohesion (Halliday and Hasan 1976).

Although a corpus annotated with various types of linguistic information is considered to be useful for different works of descriptive linguistics, applied linguistics, and language technology; the process of annotation (both manual and automatic) invariably asks for long-time involvement of trained experts with pinpointed efforts to come up with benchmark standards to be used in a uniformed manner across all language types for creation of the annotated corpora (deHaan 1984). However, anyone who wants to annotate a text will have to deal with the following two important questions (Leech 1993, Leech 2005):

(a) What kind of linguistic information should be annotated in the corpora, and

(b) How it should be annotated (manually or automatically).

For the first question, we can come up with well-defined schemes, which will allow us annotate various intralinguistic and extralinguistic information in a corpus. These schemes are related to *spoken text transcription, orthography, part-of-speech, morphology, grammar, syntax, semantics, anaphora, discourse, pragmatics, sociolinguistics* and others.

With regard to the second question, we may annotate, at the time of annotation, only one type of information in the text and ignore other types, if we understand that other types of information is not required. This, however, does not imply that other types of information are not required or possible to annotate in the corpus. We are always free to add, as and when required, other types of annotation to a corpus already annotated with one type. Therefore, we argue that an annotation scheme should be developed in such a manner that it supports various types of annotation in one or multiple layered interfaces.

Moreover, there should be no compromise with the amount of information to be annotated to a corpus. In fact, the more of information annotated to a corpus, the utility of the corpus is more enhanced, because an annotated corpus becomes more useful for varieties of

linguistic investigation and application (Grice, Leech, Weisser and Wilson 2000, Hardie 2003).

3 State-of-the-art of Corpus Annotation

Language corpora annotated at various levels and types are now available for many advanced languages like *English, Spanish, French, German, Italian, Chinese,* and *Japanese,* etc. (Leech 2005, Hunston 2002, McEnery 2003, O'Donnell 1999, Sinclair 1994, Archer and Culpeper 2003). In global perspective, the number of POS tagged corpora is much higher than other types of annotated corpora due to the following reasons.

(a) The process of POS annotation is comparatively easier than other types of annotation. Also, it can be easily applied (manually or automatically) on freely available written and spoken corpora of different forms, formats, types, and contents.
(b) Non-experts with rudimentary knowledge about morphological-cum-grammatical information of words can annotate words at part-of-speech level in a corpus.
(c) Till date, POS annotated corpora have shown greater applicational relevance than other types of annotated corpora. The POS annotated corpora are readily used in different works of descriptive linguistics, applied linguistics and language technology.
(d) The free availability of tools and software for POS annotation has worked as a catalyst for developing this type of corpora than other types.
(e) Achieving high rate of success in POS annotation is highly possible with simple trial, verification, and modification of annotation rules (Leech 2005).
(f) Other types of corpus annotation require highly specialized knowledge even for achieving a very small amount of success. People without adequate knowledge about phonetics, phonetic transcription, intonation, supra-segmental features, and other properties of speech can hardly annotate a speech corpus. Similarly, without sound knowledge in semantics, syntax, discourse, and pragmatics one may fail at every step of semantic, anaphoric and discourse annotation.

Due to such factors, the number of corpora annotated at other types is far below than the number of POS annotated corpora.

The Indian languages cut a sorry figure in case of corpus annotation (Dash and Chaudhuri 2000, Dash 2008). Till date, a few POS annotated text corpora are developed in some of the Indian languages (http://tdil-dc-in) and these are neither varied, nor large in size, nor user-friendly (Dash 2013). Moreover, tools and software for annotating Indian language corpora are not yet properly developed due to technical and motivational deficiencies. (Baker, Hardie, McEnery, Cunningham, and Gaizauskas 2002). But the most striking deficiency is the lack of properly developed text and speech corpora in the Indian languages.

Nevertheless, the present Indian scenario is rapidly changing towards a better state where corpora in a number of Indian languages with different types and formats of text annotation are increasing day-by-day. For instance, the ILCI-I tagged corpus of Indian languages contains approximately 10 million POS tagged words covering 12 Indian languages (Jha *et al.* 2011). Also, the pressing needs of Indian language technology efforts and the difficulties involved in the activities have inspired many scientists across the country to take up the challenge of corpus creation and annotation (Hardie 2003, Hardie 2005, Hardie, Koller, Rayson and Semino 2007). Therefore, it is not a difficult task to make a tentative estimation about the present state of corpus creation and annotation in Indian languages (Table 1).

Annotation Types	Availability in Indian Languages
Orthographic Annotation	Some corpora are available in Indian languages, particularly in case of transcription of spoken texts into written form
Grammatical Annotation	Available for majority of Indian languages including Hindi, Urdu, Sanskrit, Punjabi, Gujarati, Konkani, Marathi, Oriya, Assamese, Bengali, Tamil, Telugu, Malayalam etc.
Prosodic Annotation	Few Indian languages are prosodically annotated, such as Hindi and Bengali, Tamil, Telugu, etc.
Semantic Annotation	No Indian language corpus is annotated at this level
Discourse Annotation	Not available in Indian languages
Anaphoric Annotation	Not found in corpus of Indian languages

Table 1: Present state of corpus creation and annotation in Indian languages

This may also reflect on the present state of research activities in the Indian languages in this sphere of knowledge harvesting, knowledge generation, and information management.

Keeping the present state and variety of corpus annotation across the world in mind we have proposed here EA in which we try to annotate the source of words used in a piece of text of a language to identify as well as record the 'mother language' from where these words are obtained and used. This annotation is necessary because a large quantity of vocabulary of a language is actually obtained from various other languages. Moreover, the actual source of origin of words used in a language needs to be properly annotated for future linguistic works. In next two sections, we have focused on vocabulary classification of the lexical stock of a language with reference to etymology (Section 4), and designed a tagset for the purpose of EA (Section 5).

4 Vocabulary Classification

Vocabulary classification is one of the most important processes of language analysis in the area of descriptive and historical linguistics. In language technology and computational linguistics also, it has turned up as an important strategy for language-specific lexical information retrieval and knowledge representation. In the act of vocabulary classification, we propose to identify the source of origin of a word and annotate it accordingly. For instance, within a modern Bengali text corpus we have annotated the word iskul/ENG/ "school" as an English word, because although the word is a part of the present vocabulary of the Bengali language, the mother source of the word is English. Therefore, it is annotated as an English word, and not as a Bengali word. In case of hierarchical tagging it should carry tags of both the languages. Through this process, we shall be able to learn words of which ancestry are used in a language and what kind of morphophonemic changes these words have undergone in the course of naturalization in the language (Rissanen 1989).

The basic goal of this process is to capture the information of the source language of a particular word that has come to be used in another language. For instance, in a language like Bengali, it has been observed that a large part of its present vocabulary is actually derived from various other languages, such as, *Sanskrit, Arabic, Persian, English, Hindi, Portuguese, Dutch,* etc. besides having words and terms inherited from its native sources (Sen 1992, Sarkar and Basu 1994, Chaki 1996, Shaw 1999). Simple analysis of a modern Bengali text corpus has shown that most of these words are actually used in naturalized form (Dash, Datta Chowdhury, and Sarkar 2009) due to which it has become really tough to trace their actual origin or etymological ancestry. This has been the controlling factor to argue for introducing the concept of EA where, at the time corpus annotation, we are willing to assign etymological information to words with regard to their antiquity for future reference and application.

It is expected that etymological information of words should be properly tagged in a piece of text in accordance with origin of words, which may, at subsequent stages, help the language investigators know from which source languages these words have come into a language. For example, based on traditional scheme of vocabulary classification, we can classify the lexical stock of a language into three broad types:

(a) **Native stock:** This includes words inherited from 'mother language' as well as from local dialects and others. For instance, for Bengali, the words obtained from *Sanskrit, Tatsama, Tadbhaba, Deshi*, and dialects may be put into this category.

(b) **National stock:** It includes words and terms obtained from other regional and national languages. For instance, for Hindi, it covers words taken from Urdu, Punjabi, Marathi, Tamil, Telugu, Malayalam, Oriya, Bengali, etc.

(c) **Foreign stock:** It includes words borrowed from foreign languages. For instance, for Hindi, words borrowed from Arabic, Persian, English, French, German, etc. are put into this category.

Given below is an etymology-based classified list of words obtained from a Bengali text corpus to show how the vocabulary of modern Bengali is made up with words of different languages (one/two words are given from each language for reference only):

(a) Native Stock
Bengali : rāstā "road", ghar "house".
Tatsama : akṣi "eye", agni "fire".
Tadbhaba : āj "today", āṭ "eight".
Indigenous: ḍiṅgi "canoe", jhol "broth".

(b) National Stock
Hindi : kāmāi "absence", lāgātār "continuous".
Tamil : curuṭ "cigar", khokā "boy".
Santhali : kurāṭ "axe", biṛā "bundle'.

(c) Foreign Stock
English : āpil "appeal", āpel "apple".
Arabic : ārji "request", kisyā "story".
Persian : kharid "buy", cāmac "spoon".
Portuguese: ālmāri "almirah", cābi "key".
German : jār "Tsar", nātsi "Nazi".
French : ātel "intellectual', byāle "ballet".
Dutch : hartan "harten", ruitan "ruhiten".
Spanish : kamreḍ "comrade", ārmāḍā "armada".
Italian : kompāni "company", gejeṭ "gazette".
Russian : spuṭnik "sputnik", glāsnast "glasnost".
Australian: kyāṅgāru "Kangaroo".
Japanese: hārākiri "suicide", hāiku "haiku".
Chinese : cā "tea", cini "sugar".
Burmese: ghughni "curry", luṅgi "lungi".
Tibetan : iyāk "yak", lāmā "Llama".
Peruvian: kuināin, "quinine".
African : jebrā "Zebra", bhubhujelā "vuvuzela".
Hybrid : slibhhīn "sleeveless", oṣṭhogrāphy, "art of kissing".
Unknown: harpoon "harpoon".

For a language or the other, such classification scheme may be modified based on the name of the source languages from where words are inherited and borrowed. For instance, while English will include many *Scandinavian, Greek, Latin, French, German, Spanish, Italian* and other languages into its list of source languages, a South Asian language like Malayalam will include many Dravidian languages, Sanskrit, English, and other Indian languages

5 Defining EA Tagset

Since most of the living languages have directly or indirectly obtained words from other languages besides using their own stock, it is expected that at the time of EA, information about the source of words should be accurately tagged in the text corpus. Therefore, we need to have a well-defined tagset that can be uniformly applied to annotate each and every word found in the corpus. For Indic languages we can think

of using the following tagset for words coming from various languages across the world (Table 2).

No	Language	Tag
01	African	[AFR]
02	Arabic	[ARB]
03	Assamese	[ASM]
04	Australian	[AUS]
05	Bengali	[BNG]
06	Burmese	[BRM]
07	Chinese	[CHN]
08	Dialectal	[DLT]
09	Dutch	[DTH]
10	English	[ENG]
11	French	[FRN]
12	German	[GMC]
13	Hindi	[HND]
14	Hybrid	[HRB]
15	Italian	[ITL]
16	Japanese	[JPN]
17	Native	[NTV]
18	Oriya	[ORI]
19	Persian	[PRS]
20	Peruvian	[PRV]
21	Portuguese	[PRG]
22	Russian	[RSN]
23	Santhali	[SNT]
24	Spanish	[SPN]
25	Tadbhaba	[TDV]
26	Tamil	[TAM]
27	Tatsama	[TSM]
28	Telugu	[TLG]
29	Tibetan	[TBT]
30	Unknown	[UNN]

Table 2: Language-based Tagset for EA

If such tags are attached with the words in the corpus it will be easier to know the actual etymological source of words used in a language. However, it should be kept in mind that annotating such information automatically or manually with the words is not a trivial task, as it asks for sound knowledge of etymological information of words on the part of the text annotators. Therefore, only those people who are well versed with the history of origin of each word may be asked to do the said task. Also, supporting information may be retrieved from etymological dictionaries available in a language to verify as well as to authenticate the information about the origin of words before these are annotated in the corpus.

Although the tagset proposed in the Table 2 above is primarily meant to tag single-level information to the words coming from different languages, we have a future plan for encoding subsequent layers of etymological information of the words. In fact, the language tags that are proposed here can roughly indicate the source language from where a particular word is borrowed. This, however, asks for a second layer of annotation (in a hierarchical order) to capture the information of origin of a word as well as the process of derivation, alternation, and euphonic changes it might have undergone in the borrower's language with a possibility for semantic change. For instance, consider the borrowed Bengali word *māine* "monthly salary". Etymologically it is derived from the Persian word *māhiyānā* "month" (cf. Hindi, *māhinā* "month"). In this case at least the word has undergone both phonological and semantic change after it is borrowed into Bengali. This information may be tagged with the word in a manner like *māine*/PRS BNG/ to indicate etymological hierarchy of the word. In our view, this kind of hierarchical annotation may be useful in case of those **portmanteau words** where the lexical items of two different languages are combined to together to form a compound word, e.g., *sinemākhor* "cinema addict", *klāśghar* "class room", *noṭbai* "notebook", *bhoṭdātā* "voter", etc. Due to shortage of space this process is not explained here in details.

The remaining part of the paper is constructed in the following order: in Section 6, we have briefly discussed the actual process of assigning tagset to words in a sample Bengali text; in Section 7, we have presented some lexical level data and information obtained from this sample tagged corpus; and in Section 8, we have highlighted the applicational benefits of etymologically annotated corpora.

6 Process of Etymological Annotation

Annotation can be done either manually or automatically. It is, however, better to annotate a text manually for the first time so that the reliability of an annotated text is beyond question, and the text is authentically used as a trial database for development of an automatic annotation system or tool.

Kṛṣṇa/SKT/ ebār/BNG/ mādhyamik/SKT/ parīkṣā/SKT/ debe/BNG/. Kṛṣṇer/SKT/ mā/TDV/ balechen/BNG/, āmār/BNG. keṣṭā/TDV/ myāṭrik/ENG/ pāś/ENG/ karle/BNG/ moṭar/ENG/ sāikel/ENG/ kine/BNG/ debo/BNG/, kaleje/ENG/ paṛte/BNG/ ýābe/BNG/. Kṛṣṇer/SKT/ bāp/TDV/ bhuṣimāler/PRS/ kārbāri/ARB/. Tini/BNG/ balechen/BNG/, osab/BNG/ habe/BNG/ nā/BNG/. Pāś/ENG/ karle/BNG/ dokāne/PRS/ basiye/BNG/ debo/ BNG/. Jami/ARB/ jiret/ARB/ nei/BNG/, dokān/ARB/ nā/BNG/ dekhle/BNG/ khābe/TDB/ kī/TDV/ ? Kaleje/ENG/ paṛe/BNG/ ki/TDV/ cākri/PRS/ karbe/BNG/? Pāś/ENG/ karle/BNG/ cārṭe/TDV/ jāmā/ARB/, duṭo/TDV/ phatuyā/PRS/, cārṭe/TDV/ luṅgi/UNN/ kine/TDV/ debo/BNG/. Otei/BNG/ habe/BNG/. baṛa/TDV/ jor/ARB/ ekṭā/TDV/ sāikel/ENG/. Tāi/TDV/ śune/BNG/ Kṛṣṇer/SKT/ man/SKT/ khub/PRS/ khārāp/PRS/. Kṛṣṇer/SKT/ ṭhākumā/TDV/ śune/BNG/ balechen/BNG/, ore/NTV/ Keṣṭā/TDV/, bhābis/BNG/ nā/BNG/. Pāś/ENG/ karle/BNG/ tor/BNG/ ekṭā/TDV/ be/TDV/ debo/BNG/. Sukhe/SKT/ saṃsār/SKT/ karbi/BNG/ ār/BNG/ bāper/TDV/ dokān/ARB/ sāmlābi/BNG/.

Fig. 1: A sample Bengali text is annotated with etymological tagset

Now, based on the tagset defined in the earlier section, we have annotated a text manually on a trial-basis. In the diagram (Fig. 1) a sample Bengali text is

presented to show how words in the corpus are manually annotated with etymological information.

In case of automatic annotation, on the other hand, a system has to be designed, which will annotate single word units as well as multiword units in the text with appropriate etymological information. For this work the system has to be supplied with a Machine Readable Etymological Dictionary (MRED) where each and every word is marked with its relevant etymological information. Moreover, the system has to be trained in such a way that it is able to retrieve relevant etymological information from the MRED and use it to annotate the words in corpora. The process of automatic annotation may be carried out in the following algorithm made with eight steps:

Step 1: Preparation of a MRED with etymological information of each word of a language.
Step 2: Integration of the MRED with an EA system.
Step 3: Run the system of normalized digital text corpora.
Step 4: System encounters a word in the corpora.
Step 5: Matches the word with the lexical stock in the MRED.
Step 6: Extracts etymological information from the MRED.
Step 7: Annotates the word in the corpus with relevant etymological information.
Step 8: Generates the annotated output.

The process, however, may be monitored by experts when the system runs on digital corpora. When the process will run, it will encounter words of different forms and structures in corpora, such as, inflected words, non-inflected words, naturalized words, frozen words, abbreviated words, compounded words, reduplicated words, multiword strings, and hybrid words, etc. (Rayson, Archer, Baron and Smith 2006). At the initial stage, the system will annotate all single words as well as compound words (both inflected and non-inflected) used in corpora to record their source of origin. In case of ambiguity, the system will directly refer to the etymological dictionaries to dissolve confusion in proper identification of the source language of a word. If a word is left untouched in the corpora, it will be verified, validated and augmented (if needed) in the MRED. Gradually, through continuous process of modification and up-gradation the system will succeed to annotate all the words in the corpora vis-à-vis in the language.

At the initial stage we have taken only the surface level understanding of etymology which may appear inadequate in subsequent stages of text annotation. To overcome this, the decision to mark words as having specific origin may be supported with the information obtained from some authoritative etymological dictionaries available in a language by which any doubt regarding the origin of those words that travel back and forth in the course of its use in a language will be dissolved. The annotated text corpora thus developed will have many things to enrich both man and machine. In case of man, the corpora will provide a clear picture about the ration of load of words of different origins in the language. In case of machine, on the other hand, it will be easy for it to identify the major patterns of distribution of words of different etymology in the corpora, and thus, it will be able to build up useful prediction strategies on the overall patterns of occurrence of words of different origin in a language.

7 Some Findings from an EA Text Corpus

For our initial study we manually tagged words at the etymological level in a modern Bengali newspaper corpus made with 1,00000 (one lakh) words. The results obtained from this tagged corpus shows that the percentage of use of words belonging to different etymological antiquities are quite useful to shed some new lights on the present status of the language as well as on the patterns of lexical stock being used in formation of text in the language (Table 3).

Words of different Etymology	Total Words	%-age
Sanskrit (Tatsama) words	10,000	10%
Bengali words	40,000	40%
Tadbhava words	20,000	20%
English words	15,000	15%
Arabic words	07,000	07%
Persian words	06,000	06%
Other words	02,000	02%

Table 3: Percentage of words of different etymology in a Bengali newspaper text corpus

If we agree to accept the tagged newspaper corpus as a representative of the modern Bengali language, then we can, perhaps, show that (as the Table 3 displays) till date both Sanskrit (i.e., Tatsama) and the Tadbhaba words constitute a major part of the modern Bengali language besides the native Bengali vocabulary, which possesses the highest percentage of words in the language. The percentage of use of English words in the language is quite large and this is clearly reflected in the table as well as in the corpus. We have observed that the number of English words in the Bengali vocabulary is growing day-by-day as a result of new scientific and technological innovations in the western world as well as due to free global internet communication and the spread of English language and culture across international borders. On the other hand, the use of Arabic and Persian words in the language is not entirely lost, even though their percentage of use has notably decreased over the years with introduction and invasion of English into Bengali life and society. The percentage of use of words of other etymology (mostly from national and foreign stock) is quite marginal and their presence in the text does not affect much in the overall stock of the vocabulary of the language. This observation may be validated with

comparative studies of some EA diachronic corpora of a language, if available.

The Table 3 presents certain statistics on possible contribution of foreign languages to the existing Bengali vocabulary. However, the statistics is deceptive in the sense that the corpus, which is used for this study is made from newspapers texts where the information of domains and sub-domains of the text is merged for the study. But we know that the stock of words vary significantly based on domains and sub-domains from where data is obtained. For example, if the domain is science and technology, one may find more English and foreign words. On the other hand, if the domain is local news, then possibility of finding more Bengali and Sanskrit words is much higher. To verify if this argument is valid, we are planning to carry out similar statistical studies on some newspaper text corpora of different domains and sub-domains. In fact, we have planned to carry out statistical studies on a few Indian language corpora to trace differences of percentage of words in different languages and to measure the inter-annotator agreement (e.g., words that are of foreign origin but are now viewed as native stock by the language community, etc) in the EA on corpora. We also plan to carry out case studies to measure how the information of annotation at etymological level can help in different NLP activities.

In general, information elicited from the data presented in Table 3 may be used for the purpose of language planning and education and dictionary compilation. In language planning, it will give language planners an idea how the linguistic resources should be designed with clear focus on the percentage of use of lexical items in the language; in language education, teachers will definitely look at the percentage of use of words of different etymology to concentrate on vocabulary teaching at different grades; while in dictionary compilation, lexicographers will invariably take note of the percentage of use of words of different antiquities in the corpus to decide over the selection of lexical stock to be used as entry words as well as headwords in the dictionary.

8 Conclusion

There are several utilities of etymologically annotated corpora. First of all, we can get valuable information to know which words are of native origin and which words are of non-native origin. Moreover, we come to know which native words have combined with foreign words to generate new compounds or hybrid words. Similarly, we come to know which native affixes are combined with foreign words to generate new words, and what kind of morpho-phonologial alternations the foreign words have undergone in the process of nativization in the language.

Such information becomes useful in case of frequency calculation of words of various origins, language teaching, and in compilation of general and foreign word dictionaries – both in printed and digital form. Moreover, after analyzing the words structurally, we can clearly show which affixes are tagged with foreign words (or vice versa) in formation of new words in the language. In essence, EA helps to get clear cut information for all kinds of inflected word, non-inflected word, naturalized word, frozen word, compound word, reduplicated word, and other words used in corpora of a language.

In the context of Indian languages, where we come across a large number of words borrowed from neighbouring and foreign languages, identification of sources of origin of words carries tremendous relevance in lexical database generation, morphological processing, part-of-speech tagging, dictionary compilation, language description, language teaching, and spelling pattern analysis of words (Hunston 2002, Rayson, Archer, Baron and Smith 2006).

Keeping these applications in view we have proposed here a tagset for EA as well as have designed a process of marking the source(s) of origin of words used in digital language corpora. We believe that this new concept of corpus annotation will expand applicational relevance of language corpora far beyond the realms of language technology and natural language processing into many other domains and sub-domains of applied linguistics, descriptive linguistics, and their neighbouring disciplines in years to come.

References

Archer, D. and J. Culpeper. 2003. Socio-pragmatic annotation: new directions and possibilities in historical corpus linguistics. In: Wilson, A., P. Rayson and A. McEnery (Eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, Peter Lang: Frankfurt. Pp. 37-58.

Atkins, S., J. Clear, and N. Ostler. 1992. Corpus design criteria, *Literary and Linguistic Computing,* 7(1): 1-16.

Baker, P., A. Hardie, A. McEnery, H. Cunningham, and P. Gaizauskas. 2002. EMILLE: a 67-million word corpus of Indic languages: data collection, mark-up and harmonisation, *LREC 2002 Proceedings*, Pp. 819-827.

Biber, D. 1993. Representativeness in corpus design, *Literary and Linguistics Computing,* 8(4): 243-57.

Chaki, J. B. 1996. *Bangla Bhasar Byakaran* (Grammar of the Bengali Language). Kolkata: Ananda Publishers.

Dash, N. S. 2008. *Corpus Linguistics: An Introduction*, New Delhi: Pearson Education-Longman.

Dash, N. S. 2011. Extratextual (documentative) annotation in written text corpora. *Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011)* Anna University, Chennai, 16th-19th December 2011. Pp. 168-176.

Dash, N. S. 2013. Part-of-Speech (POS) Tagging in Bengali Written Text Corpus. *International Journal on Linguistics and Language Technology.* 1(1): 53-96.

Dash, N. S. and B. B. Chaudhuri. 2000. The process of designing a multidisciplinary monolingual sample corpus, *International Journal of Corpus Linguistics,* 5(2): 179-197.

Dash, N. S., P. Dutta Chowdhury and A. Sarkar. 2009. Naturalization of English words in modern Bengali: a corpus-based empirical study. *Language Forum.* 35(2): 127-142.

deHaan, P. 1984. Problem-oriented tagging of English corpus data. In: Aarts, J. and W. Meijs (eds.) *Corpus Linguistics,* Amsterdam: Rodopi, pp, 123-139.

DeRose, S. J. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics.* 14(1): 31-39.

Francis, W. N. 1980. A tagged corpus: problems and prospects. In: Greenbaum, S., G. Leech and J. Svartvik (eds.) *Studies in English Linguistics: In Honour of Randolph Quirk,* London: Longman. Pp. 192-209.

Garside, R. 1987. The CLAWS word-tagging system. In: Garside, R., G. Leech and G. Sampson (eds.) *The Computational Analysis of English: a corpus-based approach*, London: Longman. Pp. 30-41.

Greene, B. and G. Rubin. 1971. *Automatic Grammatical Tagging of English.* Technical Report, Department of Linguistics, Brown University, RI.

Grice, M., G. Leech, M. Weisser, and A. Wilson. 2000. Representation and annotation of dialogue. In: Dafydd, G., I. Mertins and R. K. Moore (eds.) *Handbook of Multimodal and Spoken Dialogue Systems*. *Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer Academic Publishers.

Halliday, MAK and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Hardie, A. 2003. Developing a tagset for automated part-of-speech tagging in Urdu. In: Archer, D., P. Rayson, A. Wilson, and T. McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL Technical Papers Volume 16. Department of Linguistics, Lancaster University.

Hardie, A. 2005. Automated part-of-speech analysis of Urdu: conceptual and technical issues. In: Yadava, Y., G. Bhattarai, R.R. Lohani, B. Prasain and K. Parajuli (eds.) *Contemporary issues in Nepalese linguistics*. Kathmandu: Linguistic Society of Nepal.

Hardie, A., V. Koller, P. Rayson, and E. Semino. 2007. Exploiting a semantic annotation tool for metaphor analysis. *Proceedings of the Corpus Linguistics 2007 conference*. Lancaster University, UK.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jha, G.N. P. Nainwani, E. Banerjee, S. Kaushik. 2011. Issues in annotating less resourced languages - the case of Hindi from Indian Languages Corpora Initiative. *Proceedings of 5th LTC*, Poznan, Poland, Nov 25-27, 2011.

Johansson, S. 1995. The encoding of spoken texts, *Computers and the Humanities,* 29(1): 149-158.

Kupiec, J. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*. 6(1): 3-15.

Leech, G. & S. Fligelstone. 1992. Computers and corpora analysis. In: Butler, C.S. (ed.) *Computers and Written Texts*, Oxford: Blackwell. Pp. 115-140.

Leech, G. 1993. Corpus annotation schemes, *Literary and Linguistic Computing*, 8(4): 275-281.

Leech, G. 2005. Adding linguistic annotation. In: Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbrow Books. Pp. 17-29

Leech, G. and Wilson, A. 1999. Guidelines and standards for tagging. In: Halteren, H.V. (Ed.) *Syntactic Word Class Tagging*. Dordrecht: Kluwer. Pp. 55-80.

Löfberg, L., D. Archer, S. Piao, P. Rayson, A.M. McEnery, K. Varantola and J. P. Juntunen. 2003. Porting an English semantic tagger to the Finnish language. In: Archer, D., P. Rayson, A. Wilson and T. McEnery (Eds.) *Proceedings of the Corpus Linguistics 2003 Conference.* UCREL technical paper number 16. UCREL, Lancaster University. Pp. 457-464.

Löfberg, L., J. P. Juntunen, A. Nykanen, K. Varantola, P. Rayson, and D. Archer. 2004. Using a semantic tagger as dictionary search tool. In: Williams, G. and S. Vessier (eds.) *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress* (Euralex 2004), Lorient, France, 6-10 July 2004. Université de Bretagne Sud. Volume I. Pp. 127-134.

Löfberg, L., S. Piao, P. Rayson, J. P. Juntunen, A. Nykänen, and K. Varantola. 2005. A semantic tagger for the Finnish language. *Proceedings of the Corpus Linguistics 2005 Conference,* July 14-17, Birmingham, UK.

McEnery, A.M. 2003. Corpus Linguistics. In: Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press, pp. 448-463.

O'Donnell, M.B. 1999. The use of annotated corpora for New Testament discourse analysis: a survey of current practice and future prospects. In: Porter, S. E. and J. T. Reed (eds.) *Discourse Analysis and the New Testament: Results and Applications*. Sheffield: Sheffield Academic Press. Pp. 71-117.

Piao, S., D. Archer, O. Mudraya, P. Rayson, R. Garside, A.M. McEnery and A. Wilson. 2006. A large semantic lexicon for corpus annotation. *Proceedings of the Corpus Linguistics 2005 Conference,* July 14-17, Birmingham, UK.

Piao, S., Rayson, P., Archer D. and A. M. McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction, *Journal of Computer Speech and Language*. 19(4): 378-397.

Rayson, P., D. Archer, A. Baron and N. Smith, 2006. Tagging historical corpora – the problem of spelling variation. *Proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491*, International Conference and Research Center for Computer Science, Schloss Dagstuhl, Wadern, Germany, December 3rd-8th 2006.

Rissanen, M. 1989. Three problems connected with the use of diachronic corpora, *International Computer Archive of Modern English Journal*, 13(1): 16-19.

Sarkar, P. and G. Basu. 1994. *Bhasa Jijnasa* (Language Queries). Kolkata: Vidyasagar Pustak Mandir.

Sen, S. 1992. *Bhashar Itivrittva* (History of Language). Kolkata: Ananda Publishers.

Shaw, R. 1999. *Sadharan Bhasabijnan O Adhunik Bangla Bhasa* (General Linguistics and Modern Bengali Language). Kolkata: Pustak Bipani.

Sinclair, J. M. 1994. Spoken language: phonetic- phonemic and prosodic annotation. In: Calzolari, N., M. Baker, and P.G. Kruyt (Eds.) *Towards a Network of European Reference Corpora.* Pisa: Giardini. Pp. 129-132.

Smith, N., S. Hoffmann and P. Rayson. 2007. Corpus tools and methods today and tomorrow: Incorporating user-defined annotations. *Proceedings of Corpus Linguistics 2007*, July 27-30, University of Birmingham, UK.

Smith, N.I. and A.M. McEnery. 2000. Inducing part-of-speech tagged lexicons from large corpora. In: Mitkov, R. and N. Nikolov (eds.) *Recent Advances in Natural Language Processing 2*, Amsterdam: John Benjamins. Pp. 21-30.

Sperberg-McQueen, C.M. and L. Burnard (eds.) 1994. *Guidelines for Electronic Text Encoding and Interchange,* Chicago and Oxford: ACH, ALLC, and AC.

Stenström, A-B. 1984. Discourse tags. In: Aarts, J. and W. Meijs (eds.) *Corpus Linguistics: Recent Developments in the use of Computer Corpora in English Language Research*. Amsterdam: Rodopi. Pp, 65-81.