Chapter 3
Corpus Editing and Text Normalization

Abstract In this chapter, we propose for applying processes like pre-editing and text standardization as some of the essential components of corpus editing and text normalization for making a text corpus ready for access across various domains of linguistics and language technology. Here, we identify some of the basic pre-editing and text standardization tasks, and we describe these works with reference to Bangla text corpus. As the name suggests, text normalization involves diverse tasks of text adjustment and standardization to improve utility of the texts stored in a corpus in manual- and machine-based applications. The methods and the strategies that we propose here to overcome the problems of text normalization are largely tilted toward written text corpus since text normalization activities relating to spoken text corpus usually invoke a new set of operations that hardly match with the normalization processes normally applied on written text corpus. The normalized version of a text not only reduces workload in subsequent utilization of a corpus but also enhances its accessibility by man and machine across all domains where language corpus has application and referential relevance.

Keywords Corpus · Pre-editing · Standardization · Normalization
Global readiness · Overlap · Term consistency · Metadata · Transliteration
Tokenization · Disambiguation · Frozen terms

3.1 Introduction

Pre-editing is a process of adjusting texts in a text corpus in order to improve the quality of the raw text data (i.e., unannotated text data) in practical applications in machine learning, language data extraction, text processing, technical terms culling, grammatical annotation, information retrieval, machine translation, etc. The resultant output is a moderately edited text corpus, which reduces the amount of workload required in post-editing of texts (Fiser et al. 2012). This implies that pre-editing is a process of adjusting texts before these texts are further processed for subsequent NLP applications. In other words, pre-editing is one such task in which text samples are

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_3

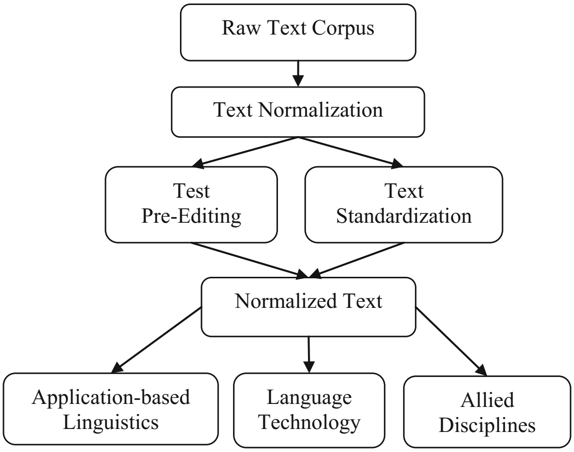

Fig. 3.1 Normalization of text corpus for cross-platform utilization

replotted in certain fixed patterns based on some pre-defined linguistic rules, such as removal of inconsistencies in expression, use of short sentences, avoidance of complex syntactic forms, simplification of ambiguous syntactic structures, redefining patterns of term consistency (Yarowsky 1996). Other pre-editing tasks may involve checking structure of words, marking multiword units, formal consistency verification of larger linguistic units like idioms, phrases, and clauses. All these tasks are required so that further text processing activities (e.g., POS tagging, chunking, parsing, lemmatization) on corpus data become simplified and trouble free (Yarowsky 1994; Xue 2003).

The text normalization process that we describe here includes two major parts: pre-editing and text standardization. Pre-editing, in simple terms, involves several text-based activities like sentence length management, typographic error elimination, punctuation inconsistency removal, header file removal, metadata management, text format simplification, lexical and syntactic ambiguity dissolution, idiomatic expression marking, orthographic style avoidance, non-textual element removal, and domain overlap prohibition (Yeasir et al. 2006). Text standardization process, on the other hand, involves activities like transliteration, grammar checking, tokenization, hyphenation, slash management, period disambiguation, white space management, frozen forms marking, emphatic particle management, indexing, cardinal number management, term usage consistency measurement. Both the parts are so deeply interlinked that occasional overlap of functions across the border is a common phenomenon in text normalization (Sproat et al. 1999, 2001). The theoretical and thematic identity of text normalization can be conceptually perceived from the diagram given below (Fig. 3.1).

3.2 Pre- and Post-editing Trade-off

There is always a trade-off between time and money spent on pre-editing and post-editing for text normalization in document processing. What is important in this enterprise is to keep in mind that if a text document is going to be translated into different target languages, it probably makes sense to spend more time on pre-editing phase than on post-editing. Also, the fact to be remembered is that pre-editing—once it is done in a source text—can solve many issues linked to post-editing (Arens 2004). Post-editing is normally done many times as the number of languages one is processing for translation or POS tagging. Pre-editing, on the other hand, is done normally once in the source text to save strangers lurking at the gate of post-editing. Therefore, the goal of pre-editing is to render the source text in such a way that the quality and output of language processing applications are upgraded. For instance, the outputs of POS tagging and machine translation (MT) will improve in terms of spelling, format of text, grammatical role of lexical items, and overall readability of text if pre-editing is carried out before the text is used as input (Chen and Liu 1992).

In order to do so, it is necessary to distinguish between those rules that improve the quality of the input text and those rules that do not affect the content of the input text. This distinction is necessary to identify the rules that are presented to the user(s) and how these rules are to be actually used for better outputs. This may involve reframing of the whole sentence in the input language at the abstract level in the sense that it should not confuse the language users in understanding the fact that the structure of the revised sentence is a restructured representation of the original input sentence (Chiang et al. 1996).

The basic argument that we advocate in this chapter is that text normalization activities offer many advantages in seamless utilization of a language corpus (Cutting et al. 1992). Therefore, it is essential to render a text corpus in such a manner that the standard of the existing NLP tools that depend on using text corpus as input is improved considerably. In order to do so, it is perhaps necessary to distinguish between the processes that improve the accessibility of texts as well as the processes that keep contents of texts intact. This is required to identify which rules are to be presented to a system and how the results of the rules may be utilized to have better outputs. This may involve reformulation of the structure of the sentence(s) in input texts at the abstract level in a manner that it should not confuse users in utilization of linguistic data and information. The ultimate goal is to create much easier text content within a corpus in respect of its readability of form, accessibility of format, and usability of content.

3.3 Pre-editing and Global Readiness

Global readiness is a process of creating and optimizing the content of a text so that the end users all over the world can grasp its meaning and intention without much effort (Abel 2011). Based on this proposition, it is fair to visualize global readiness as a multistep process of creating better text by planning, analyzing, and auditing the text for wider application in various domains of NLP. This implies that pre-editing, as a part of global readiness, makes a text ready to play a crucial role in the larger scene of language processing as well as makes a text ready for all sorts of application-based linguistic works.

Translating texts from one language to many languages—either by a man or a machine—takes more time than one language to another. That means, translating texts into multiple languages is a costly proposition on the scale of time, energy, money, effort, and efficiency. It is not that human translators or a translation system is at fault in this enterprise. In reality, both man and machine try to do the best of their ability with the source texts they receive as inputs. The problem lies with the level of complexities involved in the source texts that eventually tells upon the skill of human translators or the robustness of a machine translation system. Keeping this argument in view, in this chapter we argue for creating 'easy text' for both human translators and NLP works including machine translation.

For developing an easy text, it is necessary to focus on the features of a text relating to readability, grammar, format, and reusability. And to achieve these qualities of an input text, one has to adopt several steps of pre-editing. For example, to make corpora of the Indian languages 'global ready' we can adopt the following means as a part of the standardization of the text content:

(a) Management of scientific and technical terms used in texts,
(b) Enforcement of standard grammar rules of the language on texts,
(c) Enforcement of standard stylistic rules on texts for simplicity,
(d) Maintenance of structural consistency of texts,
(e) Elimination of unnecessary words and lexical items from texts,
(f) Shortening of unwisely sentences and segments in texts,
(g) Marking of idiomatic phrases and set expressions in texts.

Since a text often suffers from a variety of naturalization problems, it is rational to streamline and standardize a text through pre-editing to overcome the problems of text processing and computation as well as to make a text global ready for end users. In the following sections, we shall concentrate on various aspects and issues of pre-editing and normalization of texts with reference to examples and instances taken from a Bangla text corpus.

3.4 Pre-editing of Corpus

In general, there are many operations that we can do to make a corpus text global ready. Particularly in the context of a corpus text being used in NLP works, we propose to deploy the following measures on the Indian languages text corpora to make these maximally usable. These pre-editing operations may be practiced in the following areas.

3.4.1 Sentence Management

Keeping sentence length unchanged in the corpus is the optimum priority in pre-editing. There should be no compromise with regard to sentence length. Identification of each sentence as a separate syntactic unit is the most important task. This is necessary to mark and measure the length of a price of text with regard to a number of sentences included in it. Each sentence should be separated from the other if these are combined together. The punctuation marks that are normally used at the end of a sentence should be treated as legitimate sentence terminal markers.

Similarly, it is necessary to identify each segment used in a corpus. Since segments are not sentences, these forms need to be marked separately. The difference between a segment and a sentence should be clearly understood and marked accordingly so that subsequent parsing process that is to be applied to the sentences is not applied over the segments. The structural difference between a segment and a sentence is shown below with examples taken from a Bangla text corpus.

(a) Segment:

(1a) বাংলার লোকসংস্কৃতির সমাজতত্ত্ব।
Bāṅglār lokasaṃskṛtir samājtattva.
“The Sociology of folk culture of Bengal”

(1b) কৃষ্ণনগরের মৃৎশিল্প।
kṛṣṇanagarer mṛṯśilpa.
“Clay art of Krishnanagar”

(b) Sentence:

(2a) বাংলা দেশের আর একটি ঐতিহ্যবাহী প্রাচীন লোক শিল্প হচ্ছে হাতে এবং চাকায় তৈরি মৃৎপাত্র।
Bāṅglā deśer ār ekṭi aitihyabāhī prācīn lok śilpa hacche hāte ebaṃ cākāy tairi mṛṯpātra.
“Another old and hereditary folk art of Bengal is hand-made and wheel-made clay plates”

(2b) যন্ত্রানুবাদের মাধ্যমে এক ভাষার লেখ্য তথ্য ও সংবাদ অন্য ভাষায় অনুবাদ করে সারা বিশ্বের সকলের কাছে তাড়াতাড়ি পৌঁছে দেওয়া সম্ভব হবে।
ẏantrānubāder mādhyame ek bhāṣār lekhya tathya o saṃbād anya bhāṣāy anubād kare sārā biśver sakaler kāche tāṛātāṛi põuche deoyā sambhab habe.
“Through machine translation, it is possible to reach to everyone in this word quickly by translating written information and data of one language into another language”.

It is also necessary to assign unique ID for each sentence (e.g., BNG_FLT_S1990: BNG = Bangla, FLT = Folklore Text, S1990 = Sentence No.: 1990) so that each of the sentences is identified as a separate syntactic unit to be processed independently.

The long sentences which are difficult to read and comprehend should be marked for their unique syntactic structure and properties. In fact, such sentences are mostly ambiguous in nature and are often confusing in meaning. Because of these features, these are quite difficult (if not impossible) to translate into another language.

Finally, verbless sentences may also be marked with the unique flag so that at the time of POS tagging and parsing, special care is taken to address the difficulties involved in such syntactic constructions. Given below are a few Bangla verbless sentences which require additional care to find their syntactic structure as well as phrases:

(3a) এ সমস্ত তাদের অতুলনীয় কুশলতার প্রমাণ।
e samasta tāder atulanīya kuśalatār pramāṇ.
"These (are) the examples of their unparallel craftsmanship"

(3b) এটি ডোকরা শিল্পের আদি পর্বের প্রথম স্তর।
eṭi ḍokrā śilper ādi parber pratham star.
"This (is) the first phase of the early stage of Dokra Art"

(3c) এই অবস্থায় বর্তমানে ডোকরা শিল্প ও শিল্পী উভয়েই দ্রুত বিলীয়মান।
ei abasthāy bartamāne ḍokrā śilpa o śilpī ubhayei drūta bilīyamān.
"At this present stage, both Dokra Art and artist (are) fast ebbing out"

3.4.2 Typographic Error Elimination

There are several types of typographic error found within words used in a corpus, which is not normalized. With regard to typography, it is possible to classify these errors into five major types as mentioned below with some examples taken from a Bangla text corpus.

(a) **Character Omission**

Here, a particular character from a word is omitted.

হাতা 'hātā'	:	হাত 'hāt'	(ā-allograph is omitted)
ধানী 'dhānī'	:	ধান 'dhān'	(ī-allograph missed)
বাবা 'bābā'	:	ববা 'bbā'	(ā-allograph omitted)
শিশির 'śiśir'	:	শিশর 'śiśr'	(i-allograph omitted)
ইস্কুল 'iskul'	:	ইসুল 'isul'	(consonant k is omitted)

(b) **Character Addition**

In a reverse way, in some cases, a character is added to a word unknowingly.

কমল 'kamal'	:	কমলা 'kamal(ā)'	(ā-allograph is added)
পালকি 'pālki'	:	পালিকি 'pāl(i)ki'	(i-allograph is added)
গামলা 'gāmlā'	:	গামেলা 'gām(e)lā'	(e-allograph is added)
ঘরোয়া 'gharoyā'	:	ঘারোয়া 'gh(ā)royā'	(ā-allograph is added) etc.

(c) **Wrong Character Selection**

In these cases, a wrong character is used within a word in place of a right character.

চাইতে 'cāite'	:	টাইতে 'ṭāite'	('c' and 'ṭ' are closely placed character)
ছাগল 'chāgal'	:	চাগল 'cāgal'	('c' and 'ch' are assigned only one key)
আঙ্গুল 'āṅgul'	:	উঙ্গুল 'uṅgul'	('ā' is changed by 'u')
প্রমাণ 'pramāṇ'	:	প্রমাল 'pramāl'	('ṇ' is changed by 'l')
অথবা 'athabā'	:	অথমা 'athamā'	('b' is changed by 'm')
নিদর্শন 'nidarśan'	:	বিদর্শন 'bidarśan'	('n' is changed by 'b')

(d) **Character Gemination**

In this case, a particular character is doubled due to some technical reasons:

করতে 'karte'	:	কররতে 'karrte'	(r > rr)
বালক 'bālak'	:	বা◌ালক 'bāālak'	(ā > āā)
কলিকাতা 'kalikātā'	:	কললিকাতা 'kallikātā'	(l > ll)
মেয়েলি 'meyeli'	:	মে◌েয়েলি 'meeyeli'	(e > ee)
লোকটা 'lokṭā'	:	লোককটা 'lokkṭā'	(k > kk)
মহারানি 'mahārāni'	:	মহহারানি 'mahhārāni'	(h > hh)
মাতামহ 'mātāmaha'	:	মাততামহ 'māttāmaha'	(t > tt)

(e) **Character Transposition**

In this case, characters are misplaced in the order of their sequential occurrence in words. The newly formed words are, however, accepted as valid words in the language due. Therefore, this process is known as real word error(RWE) (Chaudhuri et al. 1996).

বালক 'bālak'	:	বাকল 'bākal'	(l...k > k...l)
বদল 'badal'	:	বলদ 'balad'	(d...l > l...d)
কমল 'kamal'	:	কলম 'kalam'	(m...l > l...m)
জমা 'jamā'	:	মজা 'majā'	(j...m > m...j)
কাটা 'kāṭā'	:	টাকা 'ṭākā'	(k...ṭ > ṭ...k)
কপাল 'kapāl'	:	কলাপ 'kalāp'	(p...l > l...p)
মাথা 'māthā'	:	থামা 'thāmā'	(m...th > th...m)
পাশ 'pāś'	:	শাপ 'śāp'	(p...ś > ś...p)

Some examples of non-real word errors, which are also generated through the process of character transposition, are cited below from the Bangla text corpus:

কলকাতা 'kalkātā'	:	কলতাকা 'kaltākā'
সাধারণ 'sādhāraṇ'	:	সাধাণর 'sādhāṇar'
পালিতপুত্র 'pālitaputra'	:	পাতিলপুত্র 'pātilaputra'
হাসপাতাল 'hāspātāl'	:	হাসপালাত 'hāspālāt' , etc.

3.4.3 *Punctuation Inconsistency Removal*

In principle, the proper use of punctuation marks in the text should be restored. In practicality, it is necessary to have consistent use of punctuation marks in texts. Since some of the symbols work as phrase and sentence boundary markers, their

Table 3.1 Storage of metadata within the header file of a text corpus

Metadata	<Title :: śāmba>, <Language :: Bangla>, <Genre :: Written Text>, <TC :: LIT>, <SC :: Fiction>, <TT :: Imaginative>, <ST :: Book>, <Year :: 1978>, <Edition :: First>, <Volume :: Single>, <Issue :: 0>, <Publisher :: Ananda>, <Place :: Kolkata>, <Author :: kālkuṭ>, <Gender :: Male>, <Age :: 60+>, <Nationality :: Indian>, <Words :: 5120>
Text	মরিতে চাহি না আমি সুন্দর ভূবনে। কথাটা আজ অন্য একটি কথার খেই ধরিয়ে দিল। ধরিয়ে দেওয়া খেই কথাটি অবিশ্যি বিপরীত। না-তে আছে হ্যাঁ। ভ্রমিতে চাহি আমি সুন্দর ভূবনে। <marite cāhi nā āmi sundar bhūbane. kathāṭā āj anya ekṭi kathār khei dhariye dila. dhariye deoyā khei kathāṭi abiśyi biparīt. nā-te āche hyā. bhramite cāhi āmi sundar bhūbane.>

syntactic and functional relevance cannot be ignored. It is, therefore, necessary to check if the requisite use of punctuation mark is present in the text. When two or more sentences are connected without a connector, the proper use of the period (e.g., *full stop*, *question mark*, *exclamation mark*) is absolutely necessary to mark a sentence boundary. Equally important are the proper uses of the comma and other orthographic symbols like '$, &, *' in the text because during POS tagging these are treated as 'residual text elements' and marked accordingly (e.g.,/RD_SYM/,/RD_PUNC/).

3.4.4 Metadata Management

The header file needs to be defined with a text data file in a corpus. The extratextual data and information need to be stored as the metadata in the header file for future reference. The metadata may include name, gender, nationality, and age of author, year of first publication, name of publisher, place of publication, edition used in corpus, type of text. The following table presents a list of items relating to extratextual information of a text and how this kind of information is stored in the metadata of a text file (Table 3.1).

3.4.5 Text Format Simplification

It should be understood that the standard Roman writing conventions like 'italics' and 'underlining' do not work well for many of the Indian languages scripts. In case of the Bangla text, for instance, the process of underlining may not be visually appealing due to the fact that a number of characters tend to use the space in the lower tier below the baseline. Moreover, the font design of some of the Bangla characters directly affects the usability of this kind of visual effect on texts. Similar argument stands valid in case of use of ‘italics’ in the Bangla text. Italic writing is not at all appealing for the Bangla font—in both printed and digital formats. Even in case of handwritten texts, the use of italics is the least choice, because this makes a text quite cumbersome in appearance. We can give one or two examples from a Bangla text corpus to show how such uses are very much rare in the language.

Underlined Text

(4a) যোয়ার বপনের পরে ২৫ মিমি বৃষ্টি পেলে যোয়ারের অঙ্কুর খুব ভালো হয়।
(4b) ẏoyār bapaner pare 25 mimi bṛṣṭi pele ẏoyārer aṅkur khub bhālo hay.

(5a) বাঁকুড়ার পোড়া মটির হাতি ও ঘোড়া এখন পৃথিবী বিখ্যাত।
(5b) Bãkuṛār poṛā māṭir hāti o ghoṛā ekhan pṛithibī bikhyāta.

Non-underlined Text

(6a) যোয়ার বপনের পরে ২৫ মিমি বৃষ্টি পেলে যোয়ারের অঙ্কুর খুব ভালো হয়।
(6b) ẏoyār bapaner pare 25 mimi bṛṣṭi pele ẏoyārer aṅkur khub bhālo hay.

(7a) বাঁকুড়ার পোড়া মটির হাতি ও ঘোড়া এখন পৃথিবী বিখ্যাত।
(7b) Bãkuṛār poṛā māṭir hāti o ghoṛā ekhan pṛithibī bikhyāta.

3.4.6 Ambiguity Dissolution

Ambiguity is a real challenge in text processing. The most sensible suggestion in this case is that it is always better to avoid the use of polysemous words in the text. However, in reality, this is simply impossible as a normal text will invariably have many words which are ambiguous. In this context, the possible suggestion is not to use more than one meaning or grammatical role of a word in the same sentence. But this is also not possible in a natural text since a text user never knows in which sense the word will be accepted by readers. Consider the following examples and ambiguities involved therein:

(8a) ছাত্র হিসেবে আমি আপনাকে বিশ্বাস করি।
(chātra hisebe āmi āpanāke biśvāas kari.)
Reading 1: “As a student I trust you”.
Reading 2: “I trust you as a student”.

(9a) মুর্খের মত জানতে চেও না।
(murkher mat(å) jānte ceyo nā.)
Reading 1: “Never try to know like a fool”.
Reading 2: “Never try to know the opinion of a fool”.

(10a) শুনেছি তুমি ভালো কাজ করো।
(śunechi tumi bhālo kāj karo.)
Reading 1: “I have heard that you work well”.
Reading 2: “I have heard that you do good works”.

Ambiguity is noted not only at the lexical level but at higher level also. In some cases, ambiguity is also noted in a sentence. Structural ambiguity is mostly caused due to the presence of immediately following word (W_2), which, if processed with the preceding word (W_1), may produce a meaning different from their respective independent meaning. That means, an entire sentence can be ambiguous if it is differently interpreted, as some of the examples from a Bangla text corpus show:

(11a) বাচ্চাগুলোকে সাজিয়ে-গুছিয়ে এইমাত্র খেতে বসলাম।
(bāccāguloke sājiye-guchiye eimātra khete baslām.)
1st Reading: "I just sat to eat after dressing the kids"
2nd reading: "I just sat to eat to the dressed kids"

(12a) একশ একটা ফুলের মালা দেবো।
(ekśa ekṭā phuler mālā debo.)
1st reading: "I shall give a garland made of hundred and one flowers"
2nd reading: "I shall give hundred and one flower garlands"

(13a) যেভাবে তুমি ডুবে আছো, সেভাবে আমি ডুবতে পারিনি।
(ýebhābe tumi ḍube ācho, sebhābe āmi ḍubte pārini.)
1st reading: "I cannot plunge as you do"
2nd reading: "I cannot plunge into that emotion where you are"

(14a) পশ্চিমবঙ্গ সরকারের দুগ্ধ বিক্রয় কেন্দ্র।
(paścimbaṅga sarkārer dugdha bikray kendra.)
1st reading: "Milk selling counter of WB Govt."
2nd reading: "Selling counter of WB Govt's milk"

We need to think of some methods through which it is possible to restrict the use of such words or to mark these works with specific notations at the time of pre-editing so that these can solve much confusion about word meanings among the text users in the subsequent use of texts in linguistics and language technology.

3.4.7 Idiomatic Expression Marking

All natural texts are full of set expressions, idiomatic expressions, proverbs, etc. The expressions like *kānā garur bhinna path* (a blind cow has a different path), *jale kumīr ḍāṅgāy bāgh* (crocodile in the water and tiger on the shore), *śāk diye māch ḍhākā* (to hide fish with green vegetables), *marā hātir dām lākh ṭākā* (the price of a dead elephant is one lakh rupees) are quite frequent in use in the texts. People argue that since it is not possible to eliminate idiomatic phrases from a text, it is better to reduce their use as much as it is possible (Raj et al. 2006). In our argument, this is also impossible in natural texts as people are free to use these expressions in texts as they like. It is, therefore, sensible to mark them separately at the time of pre-editing of a text by using chunking method or by some other methods considered suitable for such purposes (Fig. 3.2).

HBT2002	[[prāchīn\JJ samay\N_NN theke\PSP]]_NP [[svāsthya\N_NN]]_NP [[ebang\CC_CCD]]_CCP [[saundarya\N_NN]]_NP [[lābh\N_NN karār\V_VM_VNG janya\PSP]]_NP [[nārkelke\N_NN]]_NP [[bibhinna\JJ bhābe\RB]]_RBP [[byabahār\N_NN karā\V_VM_VNG hayeche\V_VAUX]]_VGF [[।\RD_PUNC]]_BLK

Fig. 3.2 Chunking on a sentence to mark phrase boundary

3.4.8 Orthographic Style Avoidance

It is better to use only Unicode compatible fonts like UTF8 in corpus generation. This solves many problems of text access, management, processing, and utilization. It is always better to use only one font consistently in the corpus, as the use of multiple fonts within a single text may create problems in data processing. The question of capital (upper case) or normal (lower case) fonts is irrelevant in case of texts for the Indian languages scripts, since the Indian language scripts do not follow the system of writing which the Roman script follows.

Similarly, it is better to avoid using word- or character-level styles (e.g., bold, italics, bigger shape of character, striking through word) that may force artificial display of characters in texts, as it is noted in many old and printed texts. Moreover, much care is needed in representation of conjunct characters (e.g., consonant clusters, compound characters), which are made with a combination of several consonant graphemes, vowel allographs, and diacritic symbols (e.g., *ntry, mprs, ṣṭy, pry*). This kind of font combinations may make a text look cumbersome. The task of pre-editing should take care to streamline such orthographic style variations to make a text ready for processing.

Finally, it should be noted that words made with the Roman script should be transliterated into the standard scripts of respective Indian languages, as the following examples show (Table 3. 2).

3.4.9 Non-textual Element Removal

All pictorial or visual elements are to be removed from a text corpus. Similarly, all diagrams, tables, images, graphs, flowcharts, pictures, etc., that are used in printed and digital texts should be removed from a digital text corpus. Mathematical notations, chemical formulae, geometric designs, etc., should also be removed. These elements cannot be translated or tagged in a corpus. It is better not to embed text into images as well as not to embed images into text. A pre-editing process must take care to confirm that pictorial elements can make a digital text corpus 'not-so-user-friendly'

Table 3.2 Original and revised text after orthographic consistency

No	Original text	Revised text
1	এই পার্থক্যের কতকগুলি ঘটেছে জৈবিক (biological) বা বংশগত এবং কতকগুলি সাংস্কৃতিক (cultural) কারণে।	এই পার্থক্যের কতকগুলি ঘটেছে জৈবিক (বায়োলজিক্যাল) বা বংশগত এবং কতকগুলি সাংস্কৃতিক (কালচারাল) কারণে।
	ei pārthakyer katakguli ghaṭeche jaibik (biological) bā baṃśagata ebaṃ katakguli sāṃskṛitik (cultural) kāraṇe.	ei pārthakyer katakguli ghaṭeche jaibik **(bāyolagikyāl)** bā baṃśagata ebaṃ katakguli sāṃskṛitik **(kālcārāl)** kāraṇe.
2	গিরিজনি আলোড়ন ভূ-পৃষ্ঠে অনুভূমিক আকারে (horizontally বা tangential direction) কার্য করিয়া থাকে। ইহাতে ভূ-ত্বকে কোথাও সংনমনের (compression) ফলে সংকোচনের (contraction) অথবা কোথাও টানের (tension) দরুণ প্রসারণের (extension) সৃষ্টি হয়।	গিরিজনি আলোড়ন ভূ-পৃষ্ঠে অনুভূমিক আকারে (হোরাইজন্টালি বা ট্যানজেন্টিয়াল ডিরেকশন) কার্য করিয়া থাকে। ইহাতে ভূ-ত্বকে কোথাও সংনমনের (কমপ্রেশন) ফলে সংকোচনের (কনট্র্যাকশন) অথবা কোথাও টানের (টেনশন) দরুণ প্রসারণের (এক্সটেনশন) সৃষ্টি হয়।
	girijani āloṛan bhū-pṛṣṭhe anubhūmik ākāre (horizontally bā tangential direction) kārya kariyā thāke. ihāte bhū-tvake kothāo saṃnamaner (compression) phale saṃkocaner (contraction) athabā kothāo ṭāner (tension) daruṇ prasāraṇer (extension) sṛṣṭi hay.	girijani āloṛan bhū-pṛṣṭhe anubhūmik ākāre **(horāijanṭāli bā ṭyānjenṭiāl ḍirekśan)** kārya kariyā thāke. ihāte bhū-tvake kothāo saṃnamaner **(kampreśan)** phale saṃkocaner **(kanṭryākśan)** athabā kothāo ṭāner **(ṭenśan)** daruṇ prasāraṇer **(eksṭenśan)** sṛṣṭi hay.

in text processing. Therefore, all such pictorial elements should be removed before a corpus is available for linguistics and language technology works.

3.4.10 Domain Overlap Prohibition

For better access to texts, overlapping in text or subject domains while collecting data for a corpus is not advised (Yarowsky 1994). One has to have precise idea of domains and subdomains during acquisition of language texts. For instance, for many linguistic reasons, poetic texts are removed from a corpus of prose text. Texts collected from one discipline should not be mixed up with texts of other disciplines if not specified and desired beforehand. Similarly, texts obtained from foreign languages including large quotations, statements should be removed from a monolingual corpus. The text corpus, unless otherwise defined and designed, should invariably be monolingual, domain-specific, subject-based, and synchronic (if possible).

3.5 Text Standardization

The most important argument in text standardization is that it should focus on the text to find the issues that may negatively affect the output of a text. It is, therefore, necessary to provide a correction option to improve the quality of a text (Olinsky and

Table 3.3 Words in Roman script and their transliteration in Bangla script

English word	Bangla transliteration	English word	Bangla transliteration
atabrine	ātebrin	cardo	kārḍo
chloroquine	klorokuin	coxa	kaksā
elytra	eliṭrā	femur	phimār
Galia	gyāliyā	lacinia	lyāsiniyā
malaria	myāleriyā	palpifer	pyālpiphār
palp	pyālp	paludrine	pyāluḍrin
plasmochin	plāsmokin	plate	pleṭ
staipes	sṭāipes	tarsus	ṭārsās
tegmina	ṭegminā	tibia	ṭibiyā
trochanter	ṭrokānṭār	pneumonia	niumoniyā

Black 2000). It is expected that standardization improves the accessibility of input text to a certain level so that it makes easier to operate processing methods on texts Panchapagesan et al. 2004). It should, however, be kept in mind that depending on language, text standardization rules and strategies may vary.

3.5.1 Transliteration

All technical and scientific terms written in foreign scripts should be transliterated in a text corpus. If possible, the same approach should be adopted for proper names coming from foreign languages, such as English and French names in Bangla text. The process of transliteration should be uniform across all text types in the language so that further discriminations do not arise at the time of named entity recognition or name database generation. For instance, given below is a list of English words and their transliterated forms taken from a Bangla text corpus (Table 3.3).

3.5.2 Grammar Checking

Syntactic errors are commonly found when grammatical concord between subject and predicate is lost within a sentence (Mikheev 2003). The responsibility of a corpus developer is to identify such errors, mark these properly, identify the nature of error, and correct such errors, manually or by rule-based manner. Some examples of grammar correction are presented below, for elucidation.

Wrong form	: তিনি সেখানে বসে পড়ল। **tini** sekhāne base **paṛla.**
Correct form	: তিনি সেখানে বসে পড়লেন। **tini** sekhāne base **paṛlen.** “He (hon.) sat down there”
Wrong form	: আমি তখন ক্লাস এইটে পরি। āmi takhan klās eiṭe **pari.**
Correct form	: আমি তখন ক্লাস এইটে পড়ি। āmi takhan klās eiṭe **paṛi.** “Then I was reading at class VIII”
Wrong form	: তোমার কোনো হেল্প লাগলে আমাকে বলে। **tomār** kono help lāgle āmāke **bale.**
Correct form	: তোমার কোনো হেল্প লাগলে আমাকে বলো। **tomār** kono help lāgle āmāke **balo.** “Let me know if you need any help”
Wrong form	: আমরা আসলে তাকে আমরা অন্তর্দর্শন বলি। āmrā āsale tāke **āmrā** antardarśan bali.
Correct form	: আমরা আসলে তাকে অন্তর্দর্শন বলি। āmrā āsale tāke (....) antardarśan bali. Actually, we call it insight
Wrong form	: নতুন শাসকবর্গ পুরানো অর্থনৈতিক ভিত্তি মেনে দিয়েছিল natun śāsakbarga purāno arthanaitik bhittii **mene diyechila**.
Correct form	: নতুন শাসকবর্গ পুরানো অর্থনৈতিক ভিত্তি মেনে নিয়েছিল। natun śāsakbarga purāno arthanaitik bhittii **mene niyechila.** “The new government accepted the old economic system”

3.5.3 Tokenization

A piece of text, in its raw format, is just a sequence of characters without explicit information about word and sentence boundaries. Before any further processing is done, a text needs to be segmented into words and sentences. This process is known as tokenization. Tokenization divides long character sequences into sentences and sentences into word tokens. Not only words are considered as tokens, but also numbers, punctuation marks, parentheses, and quotation marks are also treated as tokens. Given a sentence, tokenization is the task of chopping it up into small pieces called words (or shorter units, for that matter), with or without inflections. The token is an instance of a sequence of characters in a text that is grouped together as a useful semantic unit (i.e., word) for processing. In alphabetic languages, words are usually surrounded by white spaces and optionally by punctuation markers or parenthesis or quotes. These elements act as fairly reliable indicators of word or sentence boundaries (Jeffrey et al. 2002).

Tokenization is always language dependent. What is fit for a Bangla text may not be fit for a Hindi text, particularly in case of inflected verb forms. For instance, compare between the Bangla form *ýācchilām* and the Hindi form *ýā rahe the*. Both are single semantic units, but one has a single lexical unit while the other has three separate lexical units grouped together for the same purpose. Therefore, for Bangla *ýācchilām* is a single word with one token, while for Hindi *ýā rahe the* is a single word with three tokens. Given below is a Bangla sentence in its normal and tokenized form.

Normal form: Example Sentence

(8a) যখন মুক চলচ্চিত্রের প্রচলন ছিল, তখন আমরা কোনো ভাষার বন্ধনে আবদ্ধ ছিলাম না।
ýakhan mūk chalacchitrer prachalan chila, takhan āmrā kono bhāṣār bandhane ābaddha chilām nā.
"When the age of silent movie was in vogue, at that time we are not bound with boundary of any language".

Tokenized Form:

যখন (ýakhan)	মুক (mūk)
চলচ্চিত্রের (chalacchitrer)	প্রচলন (prachalan)
ছিল (chila)	তখন (takhan)
আমরা (āmrā)	কোনো (kono)
ভাষার (bhāṣār)	বন্ধনে (bandhane)
আবদ্ধ (ābaddha)	ছিলাম (chilām)
না (nā)	

3.5.4 Hyphenation

Usually, in case of a hyphenated word, hyphen carries the value of a punctuation mark and, therefore, is treated as a separate token. The main purpose of a hyphen is to glue words together. It notifies the reader that two or more elements in a sentence are linked together. Although there are rules and norms governing the use of hyphens, there are situations when we decide whether to add it or not because it can create problems while POS tagging of hyphenated words. For instance, look at the sentence *phaler ojan 70-80 grām paryanta hay* 'The weight of fruit goes up to 70-80 g'. It illustrates the fact that when a hyphen comes in between two sets of words, then a form like '70-80' is to be considered as a single-word unit and it is to be tagged as phaler\N_NN ojan\N_NN 70-80\QT_QTC grām\N_NN parýanta\PSP hay\V_VAUX .\RD_PUNC.

In this case, at least, we miss out an important piece of information of tokenized property or tokens. The string 70-80 is actually two tokens, not one. It separately conveys the idea of having something of the number between 70 and

80. Therefore, instead of keeping them together as one unit, it is better to write them as ‘70-80’. It constitutes three separate tokens: {70}, {-}, and {80}. Since {-} should be considered as a different token, the sentence should be rewritten as: *phaler ojan 70-80 grām parýanta hay*. There lies a white space between the three tokens mentioned above. In this case, the POS-tagged output will be something like the following: {phaler\N_NN ojan\N_NN 70\QT_QTC -\RD_PUNC 80\QT_QTC grām\N_NN parýanta\PSP hay\V_VAUX .\RD_PUNC}.

The same logic stands valid for the sentence *tāke ei chabir sahakārī-paricālako karā hay* ‘He is also made the assistant director of this film.’ Keeping in mind the concept of tokenization, the sentence mentioned above should be tagged as the following: {tāke\PR_PRP ei\DM_DMD chabir\N_NN sahakārī\JJ -\RD_PUNC paricālako\N_NN karā\V_VM_VNG hay\V_VAUX .\RD_PUNC}.

There are always some problems that are different from the one mentioned above. For instance, consider the following sentence: *bājir gāner rekarḍiṃer samayi gurudatta ebaṃ gitā rāy eke-aparer kāchākāchi āsen* ‘Guru Dutta and Gita Ray got closer to each other during the recording of the songs of Bazi.’ Here, the hyphenated word *eke-aparer* should be considered as a single word string made of three tokens including the hyphen in between. But again, *eke-aparer* represents the same meaning as the other form *parasparer* ‘to each other’ means. Now, the question is how to tag this string in the corpus. In our view, four possible solutions may be considered to overcome this problem:

(a) Break the string *eke-aparer* as three separate tokens as {eke} {–} and {aparer}, and tag them as {eke\PR_PRC} {–\RD_PUND} and {aparer\PR_PRC}.
(b) Since it conveys one meaning, keep it as a single token (*eke-aparer*) and tag it as a reciprocal pronoun {eke-aparer\PR_PRC}.
(c) Remove hyphen and tag the words as two separate reciprocal pronouns, such as {eke\PR_PRC} and {aparer\PR_PRC} as a whole.
(d) Remove the hyphen and tag it as a single reciprocal pronoun, such as {ekeaparer\PR_PRC} as a whole.

The decision has to be taken fast, and it entirely depends on a text annotator. In case of some complex examples such as *do-ãś mṛttikā cāṣer janya khub uapayogī* ‘Alluvium soil is best suited for farming,’ the word with a hyphen mark (i.e., *do-ãś*) constitutes a single-word unit. If we break it into two different tokens, the meaning of the word is lost. Hence, it should be tagged as a single common noun {do-ãś\N-NN} and not as an adjective and a noun. Similarly, forms like *bren-sṭem, tāntrik-tantra, cau-pāyā, spliṭ-brenoyālā, karpās-kalosām* should be tagged either as single-word units or as two-word units as these forms carry a hyphen mark in between the two formative elements.

3.5.5 Slash (/) Problem

In a written text, we sometimes come across forms like ‘9/10 din’ meaning ‘9/10 days.’ This refers to a part of time spanning over 9 or 10 days. It can be tokenized in the following two ways.

(a) If the symbol ‘/’ signifies OR function where the task can be performed in 9 or 10 days, the symbol ‘/’ can be tokenized separately and tagged as a separate punctuation mark: {9\QT_QTC} {/\RD_PUNC} {10\QT_QTC} {din\N_NN}. This is normally done according to a convention adopted in the BIS tagset.

(b) On the other hand, if form *9/10 din* quantifies something as a whole, it should not be tagged as separate units. Rather, it should be tagged as a single composite unit within the category of QT_QTC: {9/10\QT_QTC \dinN_NN}.

Let us look at some other example such as *1/3 aṃśa* ‘1/3 part.’ What should be done with this token? Can we tag it as three separate tokens because collectively they denote a measurement of something and we cannot separate the number ‘two-third’? If we do so, it will convey a different concept. We argue that there is no point in tagging as three separate tokens are here. We should take care while we select texts for the corpus, so that it will not create doubts in user’s mind if it is an OR separator or it signifies ‘a part relationship.’ We can create a uniform rule of inference for POS tagging. The POS tagging of the whole text may be as the following:

{gāch\N_NN lāgānor\V_VM_VNG samay\N_NN ālgā\JJ pātāguloke\N_NN bheṅge\V_VM_VNF mūler\N_NN ek tṛtiyāṃśa\QT_QTC keṭe\V_VM_VNF ropaṇ\N_NN karle\V_VM_VNF gāchṭā\N_NN māṭi\N_NN dhare\V_VM_VNF ney\V_VM_VF .\RD_PUNC}

Therefore, during standardization of corpus text, it is always advisable to avoid using ‘/’ whenever it signifies duration. In that case, it will be easy to tag those words.

3.5.6 Period (.) Disambiguation

In English, ‘.’ or period is considered as a punctuation mark that indicates the end of a declarative sentence or statement. In Indian languages, the same function is carried out by *pūrṇacched* ‘full stop’ (‘||’). The period is also used in Indian language texts, and in most cases, it is not used as a sentence terminal marker, but for some other functions. If a period (‘.’) appears in an Indian language text, it is mostly used to refer to an abbreviated form of nouns, e.g., *ḍ.*=*ḍāktār* ‘doctor,’ *sṭ.*=*sṭeśan* ‘station,’ *gh.*=*ghaṇṭā* ‘hour,’ *mi.*=*miniṭ* ‘minute,’ *se.*=*sekeṇḍ* ‘second.’ In all such cases, a period has one specific function, it is an indicator of the full form of the noun, hence, it should be tagged with the abbreviated form, and both of them should be considered together as a single-word unit. The problem arises when multiple nouns are abbreviated with recurrent use of period, and all the forms are meant to be put together as a single concept or expression as the following examples show:

Bangla : গতকাল বি.বি.সি থেকে দু জন লোক এসেছিল।
gatakāl bi.bi.si. theke du jan lok esechila.
English: Yesterday two people came from B.B.C.

Bangla: মাটিতে ক্যালসিয়ামের সাত পি.পি.এস. মাত্রা ফলনের জন্য আবশ্যক।
māṭite kyālsiyāmer sāt pi.pi.es. mātrā phalaner janya ābaśyak.
English: Seven P.P.S. doses of calcium are required in the soil for the production.

The question is whether we should treat *bi.bi.si.* (B.B.C.) as a unit of single token or three separate tokens. The rule holds in English that if there is a period (.) within a word, it will not be segmented; instead, it will be treated as a single unit. If this is so, then we should tag this abbreviated form as: {bi.bi.si.\N_NN}. On the other hand, the counter-argument is that the form *bi.bi.si*. 'B.B.C.' is actually made with three abbreviated forms each one of which stands for an independent word: bi. = bṛṭiś (British), bi. = braḍkāṣṭiṃ (Broadcasting) , si. = karporeśan (Corporation). It should, therefore, be treated as three separate entities and tagged accordingly: {bi.\N_NN} {bi\N_NN} {si.\N_NN}. Similarly, in a sentence like *ār. si. boṛāler janma 19-e akṭobar 1903-e ek prasiddha saṅgīt gharānār paribāre hayechila* 'R.C. Boral was born on 19th October 1903 in a highly famous family of musical tradition,' the abbreviated forms *ār. si.* should be treated as two separate words {ār.\N_NNP} and {si.\N_NNP} rather than as a single word {'ār.si.\N_NNP}, as they stand for two proper names (named entities).

3.5.7 White Space

It is necessary to remove the unnecessary white space existing between the words or tokens within a piece of text, as the following examples show.

সহজ সাধ্য (sahaj sādhya)	>	সহজসাধ্য (sahajsādhya)	"easy"
বীজ গুলি (bīj guli)	>	বীজগুলি (bījguli)	"seeds"
পেরে ছিল (pere chila)	>	পেরেছিল (perechila)	"had done"
কথা গুলো (kathā gulo)	>	কথাগুলো (kathāgulo)	"the words"
দিয়ে ছিলেন (diye chilen)	>	দিয়েছিলেন (diyechilen)	"had given"
নরেন কে (naren ke)	>	নরেনকে (narenke)	"to Naren"
মেয়ে দের (meye der)	>	মেয়েদের (meyeder)	"to girls",
উত্তর প্রদেশ (uttar pradeś)	>	উত্তরপ্রদেশ (uttarpradeś)	"Uttar Pradesh" etc.

Since these are single-word units, there is no need to give space in between the two formative parts. On the contrary, it is equally necessary to give proper space between the words where it is needed, e.g., {upakārī.er>upakārī . er} or {ṭāṭkā,lobhanīya>ṭāṭkā, lobhanīya}, {bapankarle>bapan karle}. Here, the two words *ṭāṭkā* and *lobhanīya* are clubbed together, but there should be a space between them because they are two separate words with two different meanings. Therefore, they should be written separately as *ṭāṭkā* ‘fresh’ and *lobhanīya* ‘attractive.’ This implies that typing error with regard to white space should be carefully eliminated from a text corpus. After splitting, these words will stand as independent lexical items with different grammatical functions and meanings, and subsequently these will be tagged under different parts-of-speech.

3.5.8 Emphatic Particles

This is another important text standardization process where emphatic particles need to be properly attached to words. In the existing style of writing in Bangla (and in other Indian languages), emphatic particles are the part of the preceding words and, therefore, should never be written separately. If these are written separately, these should be considered as conjuncts and not as emphatic particles. In the following Bangla examples, particles *-o* and *-i* do not work as conjuncts, but rather work as emphatic particles, and therefore, they should be tagged with their immediately previous words, as shown below:

লাগাইয়া ও ইহা টানা যায়	:	লাগাইয়াও ইহা টানা যায়
{lāgāiyā} {o} ihā ṭānā ẏāy	:	{lāgāiyāo} ihā ṭānā ẏāy
যন্ত্রের সাহায্যে ও জমিতে	:	যন্ত্রের সাহায্যেও জমিতে
ẏantrer {sāhāẏye} {o} jamite	:	ẏantrer {sāhāẏyeo} jamite
এই পদ্ধতিতে ই হস্তচালিত	:	এই পদ্ধতিতেই হস্তচালিত
ei {paddhatite} {i} hastacālita	:	ei {paddhatitei} hastacālita

In the reverse process, it is noted that the conjunct ‘o’ is sometimes tagged with the previous word as an emphatic particle. This is also a wrong representation of words. In these cases, the conjunct should be detached from the previous word and should be used as a separate lexical item in the text, as the following examples show:

ব্যবধানে ওগভীরতায়	:	ব্যবধানে ও গভীরতায়
byabadhāne {ogabhīratāy}	:	byabadhāne {o} {gabhīratāy}
রামও সীতা	:	রাম ও সীতা
{Rāmo} Sītā	:	{Rām} {o} Sītā
অঙ্গ, বঙ্গও কলিঙ্গ	:	অঙ্গ, বঙ্গ ও কলিঙ্গ
aṅga, {baṅgao} kaliṅga	:	aṅga, {baṅga} {o} kaliṅga

In the above examples, the character ‘o’ acts as a conjunct. So, it should be separated from its preceding words as well as succeeding words. Since it is a conjunct, it has its own syntactic-cum-semantic function, and thus it should be treated accordingly in the text.

3.5.9 Frozen Terms

In the present Bangla text corpus, there are some forms like HNO_3, H_2SO_3, H_2SO_4, H_3PO_4 which are normally tagged as *Frozen Forms* as these are universally acknowledged as iconic in form and meaning. In the act of text normalization and processing, these forms should remain same for any text of any language. Within this category, we also have mathematical signs (e.g., $\times$, $\div$, +, $-$, %, /,$<$,$>$,$=$, $\sum$, Ω), currency symbols (e.g., $, £, ¥, ₹, €), and some specific text symbols (e.g., #, &, @, ©, §, ®, ¢) which should not change in form. They fall into the category of symbol and should be treated in a formal way.

When we come across a character string something like '70%' in Indian language text corpus, we first change the Roman numeral into Indian language numeral and keep the percentage sign (%) separated from the number because this sign carries specific symbolic function and tag. Therefore, the string is taken up as two different tokens and not as a single one.

In text standardization process, we come across another problem relating to symbols such as this: 15:15:15. Since this denotes a relationship of ratio, the symbol ':' carries mathematical information. So, we need to write it in the following format: '15: 15: 15' keeping a space between the digit and the symbol. Similarly, we also come across a string like '6:30:44,' which denotes time indicating hour, minute, and second—all tagged together with the use of the colon (:) between the characters. In this case, also, we have to break the string into three separate units like, '6': '30': '44' and specify that each unit separated by a colon is actually indicating a separate lexical unit with separate meaning and function (though with identical part-of-speech). Alternatively, if the text standardization task is not so rigorous and lexical-bound, one can, for simple comprehension, keep the entire string as an unbroken unit and tag accordingly {\6:30:45\N_NN}.

3.5.10 Indexing

It is noted that most of the Indian languages texts use the Roman numerals in place of standard Indic script numerals. This creates a problem in text processing. To overcome this, we suggest that all the Roman numeral characters should be converted into Indian numeral characters or vice versa. Similarly, all English alphabets should be converted into Indian alphabets at the time of enumeration, for example: (a) = (k), (b) = (kh), (c) = (g), (d) = (gh), (e) = (ṅ).

When we come across digits such as (1) or letters such as (k) within a bracket, they should be treated as single tokens, while the brackets encircling them should be treated as separate symbols. Therefore, it will be better if such strings are marked in the following manners:

(1): {(\RD_PUNC, 1\QT_QTC,)\RD_PUNC}
(k): {(\RD_PUNC, k\QT_QTC,)\RD_PUNC}.

3.6 Conclusion

In this chapter, we suggest that corpus editing and text normalization are necessary for the text corpora of the Indian languages because they offer many advantages in seamless utilization of language texts stored in the corpora (Habert et al. 1998). The goal is to render the source text in such a manner that the existing standard of activities of language technology is considerably improved so that the problems of spelling, format of text, grammatical roles of words, and overall readability of a text, etc., do not create serious hurdles in use of language corpora (Huang et al. 2007).

In order to do so, we need to distinguish between those rules that improve the quality of the input text and those that do not affect the quality of a text in corpus. This distinction should be maintained as it is necessary to identify which rules are presented to the user(s) and how the result of the rules can be used by the text users to have better application outputs. This may even involve reformulation of the whole sentence(s) in the input language at the abstract level in the sense that it should not confuse the text users in the use of language data and texts. The ultimate goal is to create a much easier content within a corpus in respect of its readability of form, accessibility of format, and reusability of content.

References

Abel, S. 2011. Ready for the World: Is Your Content Strategy Global Ready? Blog on 7 April 2011 at: http://thecontentwrangler.com/2011/04/07/ready-for-the-world-is-your-content-strategy-global-ready/.

Arens, R. 2004. A Preliminary Look into the Use of Named Entity Information for Bioscience Text Tokenization. In *Proceedings of the Student Research Workshop (HLT-SRWS'04), HLT-NAACL-2004*, 37–42. PA, USA: Association for Computational Linguistics Stroudsburg.

Chaudhuri, B.B., and U. Pal. 1996. Non-word error detection and correction of an inflectional Indian language. In *Symposium on Machine Aids for Translation and Communication (SMATAC-96)*, New Delhi, April 11–12, 1996 (Hand out).

Chen, K.J., and S.H. Liu. 1992. Word Identification for Mandarin Chinese Sentences. In *Proceedings of the 14th Conference on Computational Linguistics*, 101–107. France.

Chiang, T.H., J.S. Chang, M.Y. Lin, and K.Y. Su. 1996. Statistical Word Segmentation. *Journal of Chinese Linguistics.* 9: 147–173.

Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun. 1992. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 133–140.

Fiser, D., N. Ljubesic, and O. Kubelka. 2012. Addressing Polysemy in Bilingual Lexicon Extraction From Comparable Corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012),* Istanbul, Turkey.

Habert, B., G. Adda, M. Adda-Decker, P. Boula de Mareuil, S. Ferrari, O. Ferret, G. Illouz, and P. Paroubek. 1998. Towards Tokenization Evaluation. In *Proceedings of LREC-98*, 427–431.

Huang, C.R., P. Simon, S.K. Hsieh, and Prevot, L. 2007. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, 69–72. Prague.

Jeffrey, T.C., H. Scuhtze, and R.B. Altman. 2002. Creating an Online Dictionary of Abbreviations from MEDLINE. *Journal of American Medical Informatics Association* 9 (6): 612–620.

Mikheev, A. 2003. Text Segmentation. In *The Oxford Handbook of Computational Linguistics*, ed. R. Mitkov, 201–218. New York: Oxford University Press, Inc.

Olinsky, C., and A. Black. 2000. Non-Standard Word and Homograph Resolution for Asian Language Text Analysis. In *Proceedings of the ICSLP-2000,* Beijing, China, (available: www.cs.cmu.edu/~awb/papers/ICSLP2000_usi.pdf).

Panchapagesan, K., P.P. Talukdar, N.S. Krishna, K. Bali, A.G. Ramakrishnan. 2004. Hindi Text Normalization. In *Presented at the 5th International Conference on Knowledge Based Computer Systems (KBCS),* Hyderabad, India, 19–22 December 2004. (www.cis.upenn.edu/~partha/papers/KBCS04_HPL-1.pdf).

Raj, A., T. Sarkar, S.C. Pammi, S. Yuvaraj, M. Bansal, K. Prahallad, and A. Black. 2006. Text Processing for Text-to-Speech Systems in Indian Languages. In *Proceedings of the ISCASSW6*, 188–193. Bonn, Germany, (www.cs.cmu.edu/~awb/papers/ssw6/ssw6_188.pdf).

Sproat R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 1999. Normalization of Non-Standard Words: WS'99 Final Report. In *Proceedings of the CLSP Summer Workshop,* Johns Hopkins University, (Available: www.clsp.jhu.edu/ws99/projects/normal).

Sproat, R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of Non-Standard Words. *Computer Speech and Language* 15 (3): 287–333.

Xue, N. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing.* 8 (1): 29–48.

Yarowsky, D. 1994. Homograph Disambiguation in Text-to-Speech Synthesis. In *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*, 244–247. New Paltz, NY.

Yarowsky, D. 1996. Homograph Disambiguation in Text-to-Speech Synthesis. In *Progress in Speech Synthesis*, ed. J.V. Santen, R. Sproat, J. Olive, and J. Hirschberg, 157–172. New York: Springer.

Yeasir, K.M., A. Majumder, M.Z. Islam, N. UzZaman, and M. Khan. 2006. Analysis of and Observations from a Bangla News Corpus. In *Proceedings of the 9th International Conference on Computer and Information Technology (ICCIT-2006),* Dhaka, Bangladesh.

Web Links

http://www.worldcat.org/title/coling-2002.
https://en.wikipedia.org/wiki/Text_normalization.
https://msdn.microsoft.com/en-us/library/ms717050(v=vs.85).aspx.
http://www.cslu.ogi.edu/~sproatr/Courses/TextNorm/.