International Journal of Applied Linguistics & English Literature
E-ISSN: 2200-3452 & P-ISSN: 2200-3592
www.ijalel.aiac.org.au

Looking into the Operational Modalities Adopted in Some of the POS Tagging Tools in Identification of Contextual Part-of-Speech of Words in Texts

Kesavan Vadakalur Elumalai[1]*, Niladri Sekhar Das[2], Mufleh Salem M. Alqahtani[1], Anas Maktabi[1]

[1]*Department of English Language and Literature King Saud University, Riyadh, Kingdom of Saudi Arabia*
[2]*Linguistic Research Unit Indian Statistical Institute, Kolkata, India*

Corresponding Author: Kesavan Vadakalur Elumalai, E-mail: ekesavan@ksu.edu.sa

ARTICLE INFO

Article history
Received: August 18, 2019
Accepted: October 17, 2019
Published: November 30, 2019
Volume: 8 Issue: 6
Advance access: November 2019

Conflicts of interest: None
Funding: None

ABSTRACT

Part-of-speech (POS) tagging is an indispensable method of text processing. The main aim is to assign part-of-speech to words after considering their actual contextual syntactic-cum-semantic roles in a piece of text where they occur (Siemund & Claridge 1997). This is a useful strategy in language processing, language technology, machine learning, machine translation, and computational linguistics as it generates a kind of output that enables a system to work with natural language texts with greater accuracy and success. Part-of-speech tagging is also known as 'grammatical annotation' and 'word category disambiguation' in some area of linguistics where analysis of form and function of words are important avenues for better comprehension and application of texts. Since the primary task of POS tagging involves a process of assigning a tag to each word, manually or automatically, in a piece of natural language text, it has to pay adequate attention to the contexts where words are used. This is a tough challenge for a system as it normally fails to know how word carries specific linguistic information in a text and what kind of larger syntactic frames it requires for its operation. The present paper takes up this issue into consideration and tries to critically explore how some of the well-known POS tagging systems are capable of handling this kind of challenge and if these POS tagging systems are at all successful in assigning appropriate POS tags to words without accessing information from extratextual domains. The novelty of the paper lies in its attempt for looking into some of the POS tagging schemes proposed so far to see if the systems are actually successful in dealing with the complexities involved in tagging words in texts. It also checks if the performance of these systems is better than manual POS tagging and verifies if information and insights gathered from such enterprises are at all useful for enhancing our understanding about identity and function of words used in texts. All these are addressed in this paper with reference to some of the POS taggers available to us. Moreover, the paper tries to see how a POS tagged text is useful in various applications thereby creating a sense of awareness about multifunctionality of tagged texts among language users.

Key words: Annotation, Tagging, Part-of-Speech, Morphology, Syntax, Semantics, Contexts

INTRODUCTION

An electronically developed corpus (i.e., digital language database), after it is annotated at the part-of-speech level, becomes useful for various works of language analysis, processing, application and reference in language technology, applied linguistics, translation, dictionary compilation, language teaching and description (Sinclair 2004). The process of POS tagging is normally carried out on a digital version of a corpus (manually or automatically) using a set of predefined tagsets, which are developed separately to assign part-of-speech to words. In general, a POS tagset normally includes information about linguistic properties and functions of words and terms that are used in a piece of text (Biber *et al.* 1994). The task of assigning part-of-speech to words, although it appears to be straightforward, simple and one dimensional, is functionally embedded with many theoretical and technical challenges. One of the tough challenges is the identification of lexico-semantic identity and syntactic-grammatical function of a word based on which its part-of-speech is determined. Another challenge is the invocation of the process that includes defining the basic hierarchical modalities of tag assignment and designing a rule-based schema for automated tag assignment to words. These issues ask for application of a synchronized strategy designed with the proper combination of linguistic and extralinguistic knowledge. It also requires computational expertise for achieving maximum precision with a minimum enterprise.

Published by Australian International Academic Centre PTY.LTD.

http://dx.doi.org/10.7575/aiac.ijalel.v.8n.6p.92

In natural language processing and language engineering, the process of POS tagging is also identified as *Grammatical Annotation* the primary goal of which is to disambiguate words at the grammatical level and assign them to particular lexical categories (Osselton 1984, Santorini 1990). Because of this unique goal, POS tagging is known as *Word Category Disambiguation*, which in essence, involves the process of marking up words in a corpus as corresponding to particular parts-of-speech, based on forms, functions, and contexts of words (i.e., relation of a word with adjacent words) within larger syntactic frames like phrases and sentences (Rayson *et al.* 2007).

The process of POS tagging is a complicated and error-prone process. Even then, we cannot ignore it because research and development work in many areas of linguistics cannot move further just with a list of words without any information about their grammatical behavior in usage-based contexts (Mueller 2005). In descriptive and applied linguistics, for instance, POS tagging of words is necessary because we find that words are able to represent different parts-of-speech in different contexts. The information about the parts-of-speech of words that we find in traditional grammars and dictionaries is fuzzy, implicit, and indeterminate (Kytö & Rissanen 1993).

This paper has a humble goal. It evaluates how some of the POS tagging systems are able to serve the purposes for which these are designed. It also wants to examine how POS tagging is necessary for linguistics and sister domains. To achieve this goal, it tries to understand the basic concept of POS tagging of a text (Section 2); proposes for expanding the list of part-of-speech of a language to address requirements of a modern text which contains a large amount of texts filled with code-switching and code-mixing (Section 3); describes the stages involved in POS tagging (Section 4); defines the CLAWS POS tagging system (Section 5); describes briefly the method of Hidden Markov Model (Section 6); tries to understand the methodology used in Dynamic Programming Algorithm (Section 7); refers to the model used in Brill POS tagger (Section 8); explores the method adopted in TnT POS tagger (Section 9); reports about the primary findings with some discussions (Section 10), and finally, identifies utility of a POS tagged corpus in natural language processing, language technology, machine learning, applied linguistics, and language description (Section 11).

WHAT IS PART-OF-SPEECH (POS) TAGGING?

Theoretically and applicationally, part-of-speech (POS) tagging is a process of assigning the appropriate part-of-speech tag to a word used in a piece of text after the word is passed through the stages of morphological analysis and grammatical interpretation (Garside 1995). Generally, a set of specially designed codes, which are known as 'tags' and which carry word-specific grammatical information, are assigned to the words to indicate their parts-of-speech with regard to their functions indicated in the text (Leech 1997). In many cases, well-defined sets of linguistic rules are applied to identify part-of-speech of words as well as to assign POS tags to words to determine their lexico-semantic identity and syntactic-grammatical functions in a piece of text (Leech *et al.* 2001). The immediate advantages of POS tagging in a text are realized at five levels:

(a) **Orthographic:** Helps to draw a distinction among homographic forms used in a text.
(b) **Morphological:** Allows analyzing morphological properties noted at the surface forms of words.
(c) **Syntactic:** Allows identifying syntactic-grammatical functions of words.
(d) **Lexical:** Allows assigning appropriate part-of-speech value to words.
(e) **Semantic:** Helps to make distinctions in semantic roles of words.

The POS tagging is the commonest form of text annotation. It is considered as the very first stage of a more comprehensive process where multiword expressions (e.g., *compound words, reduplicated forms, idiomatic expressions, proverbial expressions, set phrases*) are to be assigned with chunking markers leading to the eventual assignment of phrase markers to each of the sentences used in a text. Although the use of POS tags on a text makes a text difficult to read and comprehend for human beings, it becomes maximally suitable for providing linguistic information needed by a computer system for differentiating between words used in different parts-of-speech (Leech & Eyes 1993). Moreover, from an application point of view, POS tagging is a useful technique as it increases specificity in data retrieval from a corpus and provides basic grammatical information about words required in semantic annotation, parsing, dictionary compilation, grammar writing, language teaching, and language planning (Piao *et al.* 2004).

In a simplified manner, the process of POS tagging on words in a piece of text is normally carried out through the following ten steps:

(a) Generation of a text in digital form for application
(b) Normalization and getting a text ready for POS tagging
(c) Identification of words within a piece of text
(d) Identification of their orthographic forms and appearances
(e) Analysis of their morphological structures and formation
(f) Identification of their syntactic (grammatical) functions in the sentence
(g) Determination of their grammatical roles and parts-of-speech
(h) Identification of their semantic roles in the sentences
(i) Assignment of POS tags following an accepted POS tagset
(j) Final verification and validation of the tags assigned to words.

All these works are carried out either manually or automatically based on the level of proficiency of a system or the human experts engaged in the process. While in advanced languages, it is mostly done through an unsupervised or semi-supervised manner, in less advanced languages, it is normally done through a supervised manner or simply manually.

EXPANDING THE LIST OF PART-OF-SPEECH

The feature that words vary in part-of-speech is not a new thing to a natural language. This feature is noted in all the living natural languages of the world – be it an advanced language like English or an endangered language like Birhor. It is typically noted that a large number of words in a natural language are ambiguous in form, meaning, and part-of-speech (Rayson *et al.* 2005). For instance, in English the word *sound* can have a different meaning, and part-of-speech like the followings based on the context of its use in a text:

(1) The *sound* of the music is very soothing.
(2) He has taken a *sound* decision.
(3) In this context, she *sounds* rational.

Quite clearly we can see that the word *sound* is actually referring to three different meanings and parts-of-speech in three different sentences given above. When we try to perform POS tagging on the word *sound* in the sentences above, we claim that it is used as a noun (NN) in the sentence (1), as an adjective (ADJ) in the sentence (2), and as a finite verb (FV) in the sentence (3). The knowledge that helps us to formulate this claim is actually derived from the sentential contexts where the word is used (Archer *et al.* 2003). A speaker who has a certain level of mastery on the language not only identifies three different identities of the word (noun, adjective, and finite verb) but also performs necessary grammatical and semantic analysis of the word based on linguistic rules and grammar of a language. Now, while a system is trained to tag words automatically like a human being in a text, it needs elaborate linguistic rules and grammatical conditions to be predefined and presented to it in a programmatic manner so that it can perform the task of identifying parts-of-speech of words in the text.

It is normally argued that the list of parts-of-speech presented in grammars and dictionaries are enough for a language to tag words in a text (Archer & Culpeper 2003). In reality, however, we find that there are more categories and sub-categories in a text which are not considered in the list of parts-of-speech recorded in grammars and dictionaries. For instance, in a language like Bengali, in standard grammars, it is recorded that the language has only 8 parts-of-speech, namely, Noun, Pronoun, Adjective, Adverb, Finite Verb, Non-finite Verb, Postposition, and Indeclinable. It is also recorded that only these parts-of-speech are required for the language and once we learn these categories, we shall have no problem in identifying part-of-speech of words used in it. In actuality, however, we find that there are some text categories, such as *demonstratives, infinitives, gerunds, conjunctions, enclitics, punctuations, quantifiers, particles, and emphasizers* which are not included in the list. We have to identify these categories, analyze their forms, decipher their grammatical roles, and understand their semantic functions if we want to develop a good POS tagger for the language.

We also need to think of incorporating part-of-speech from other languages that are not found in a language. This happens quite often in case of those texts where code-switching and code-mixing are common practices in regular acts of text generation. For instance, in Bengali, there is no part-of-speech like *determiner* (Det) and *preposition* (Prep). However, while we analyze modern Bengali text samples, we find that English determiners and prepositions have come into use in modern Bengali texts quite frequently. Therefore, it becomes necessary for us to include tags for determiner and preposition in the tagset designed for the Bengali POS tagging. Such decisions have to be distinctly spelled out if we want to design a POS tagging scheme for words used in a language.

STAGES OF POS TAGGING

Following the steps stated above (Section 2), the process of POS tagging is carried out on a piece of text at three separate stages as the followings (Dash 2011):

(a) **Stage 1:** Manual or automatic pre-editing of a corpus text
(b) **Stage 2:** Manual or automatic tag assignment to words
(c) **Stage 3:** Manual post-editing of the tagged text database.

At the pre-editing stage, a language text database is converted into a suitable digital format for carrying out a POS tagging programme. At this stage, the entire text database is checked to verify if there is any typographical mistake or orthographical error of any type within the text database. If any error is noted, it is manually or automatically corrected in accordance with the physical source text before the digital version of the text is put to POS tagging (Dash 2011). Moreover, if required, selected text databases may pass through the stages of text normalization and tokenization to make the database maximally suitable for POS tagging, which can be done either manually or automatically (Archer *et al.* 2004).

The tag assignment stage begins with the assignment of just one and only one POS tag to each word used in a sentence. At the initial stage, before the whole process is put to automation, this is done manually as a trial basis when a specific syntactic-grammatical function of the words in a sentence is taken into consideration. To achieve a higher level of accuracy at this stage, human annotators may use structured knowledge texts like dictionaries and grammars where words are previously assigned with possible parts-of-speech for reference purposes. Such resources are open-ended in the sense that these are updated with the addition of new words obtained from various sources of language use. Moreover, these words are supported by linguistic information of various kinds to make these lexical data usable in all kinds of linguistic applications of a language. Moreover, to deal with those newly found words, which are not available in previously designed lexical databases, human annotators adopt different methods such as lists of common affixes and case markers with their possible parts-of-speech value. This helps a human annotator as well as a system to achieve greater accuracy in POS tagging (Biber *et al.* 1998: 258-259).

At the post-editing stage, the tagged text is checked, either manually or automatically, to see if the words are rightly tagged. It also checks if any error is made in the POS tag assignment and if so what kind of error is generated. Usually, the tagged text is checked manually. However, in case of a large corpus where manual verification of the entire database is a time-consuming, tedious and error-prone task,

a system can do it. In that case, a system normally devises a probability matrix from the tagged corpus to deal with problems of non-tagging, ambiguous tagging, and dubious tagging (Leech *et al.* 1983). The matrix helps the system to specify transition probabilities underlying between the adjacent tags. For example, in a language like Bengali, if a given word is tagged as a noun ($W_{\_NN}$), the probability of its immediately preceding word to be an adjective ($W_{\_ADJ}$) or a noun ($W_{\_NN}$) is much higher than its probability to be a verb ($W_{\_FV}$) or a pronoun ($W_{\_PN}$).

Usually, a human being, who is engaged in assigning POS tags to words manually, can do the work successfully provided he is well acquainted with the morphology and grammar of a language. On the other hand, a computer system can also do this task successfully, if the system is properly trained with an adequate amount of linguistic data, information and rules for tag assignment as well as it is monitored to do the work with less percentage of errors. That means a system designer who is engaged in designing a system for automatic POS tag assignment to words should be properly equipped with adequate linguistic and grammatical knowledge of a language so that he is able to develop a system that can be fast, robust, and accurate in assigning correct POS tag to words used in a text (Kupiec 1992). However, before the task of POS tagging is executed on a text corpus, either manually or automatically, there is an urgent need for standard POS tagset, which is hierarchical, well-defined, and usable for a language. Moreover, there is a question of acceptability which implies that a tagset is accepted, adopted, and used in a uniformed manner by one and all for tagging texts of all kinds.

In the following five sections, we shall briefly focus on the basic principles and strategies that are adopted for developing the five most frequently used POS tagging methods in English. We shall also see how these systems excel or fail to address the requirements of languages like Bengali which belongs to the group of the least digitally resourced languages. Such a short analysis directly serves one of the important questions of the paper which wants to know if these POS tagging tools are at all useful for those languages which are not much digitally processed and which follow different orthographic, linguistic, and textual representation. Observations made in these sections will give ideas about the present status of the systems as well as help to gather insights for developing the POS tagging system for those languages which still lack it. However, before we go into some details of operation of the five tagging mmethods, we present below a diagram (Figure 1) which tentatively refers to the POS tagging methods developed so far. This will give some ideas to know the types of POS tagger available to us as well as select the most appropriate one for a particular language in accordance with its forms, structure and composition.

THE CLAWS POS TAGGER

The *Constituent Likelihood Automatic Word-tagging System* (CLAWS) for POS tagging for English texts is first developed at *Lancaster University*, UK. From the early 1980s, this system has been continuously used and revised several times to give its final shape (Fligelstone *et al. 1996*). The fourthrevised version of this tagger (CLAWS4) has been used to POS tag more than a hundred million words of the *British National Corpus* with an appreciable rate of accuracy. This system has consistently achieved 96 to 97% accuracy in POS tagging even though the precise degree of accuracy varied

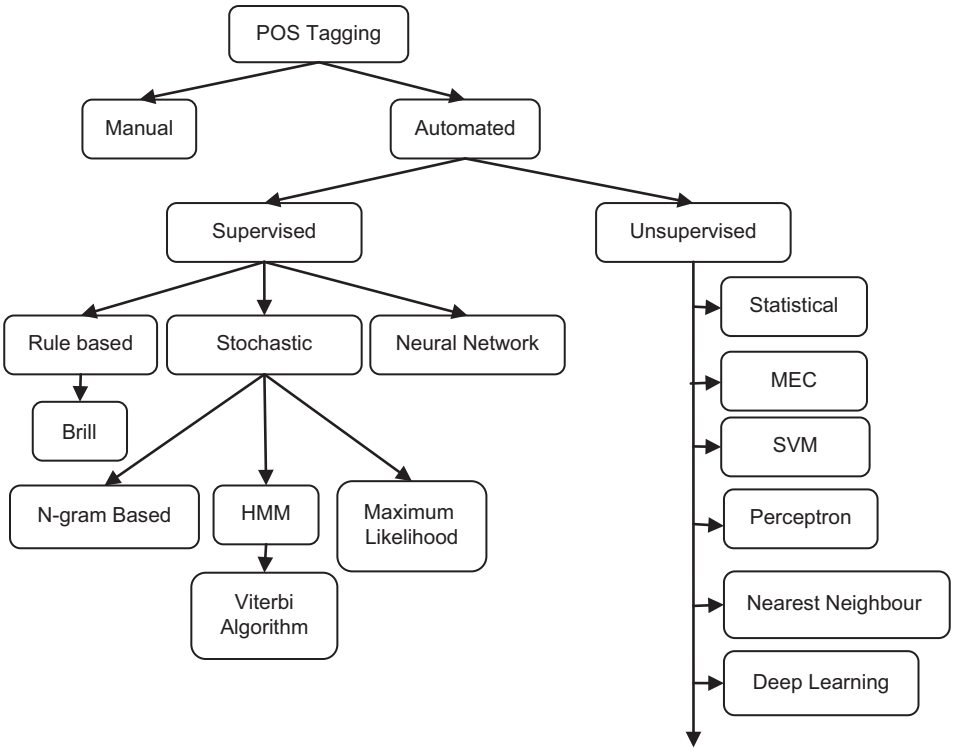

Figure 1. Tentative classification of POS tagging methods developed so far

based on the primary type of an English text (Garside & Smith 1997). Experiments carried out on the major part-of-speech categories shows that the system has an error rate of only 1.5%, and nearly 3.3% of words are ambiguous. In most cases, ambiguities of these words have remained unresolved. A template tagger – a tool of rule-based formalism – is built into the CLAWS4 version to act as a post-processor of a tagging process that can minimize the level of ambiguities (Garside 1996). The template tagger is primarily designed based on the information derived from manual analysis of tagged corpora as well as from the knowledge after analysis of frequent errors created by the CLAWS4. The implementation of the template tagger has drastically improved the tagging accuracy in the resulting corpus database (Fligelstone *et al.* 1997).

Several tagsets have been proposed and used in CLAWS. These tagsets have also been modified time and again to make these maximally appropriate for modern English texts. The tagset of CLAWS1 included 132 basic tags for words, many of which were identical in form and application to those tags that are already used in the *Brown Corpus* (Garside & Smith 1997). The revised version of the tagset that has been used in CLAWS2 has further been enlarged to include 166 tags. The logic behind the expansion of the tagset was the motive for capturing finer distinctions in grammatical functions of words in texts. However, since such an elaborate tagset, in practicality, created problems in assigning tags to words, the number is reduced to a manageable level for general purposes. Thus the revised tagset that is used for the texts included in the *British National Corpus* contained only 60 tags. This tagset is designed primarily for handling much larger quantities of databases than the databases which are specially designed for research-specific purposes. On the other hand, a sample database of the *British National Corpus* which is shared and disseminated for academic purposes contains more than 160 tags. The revised standard tagset of CLAWS is the C7_Tagset, which is advanced, elaborate and exhaustive with addition of tags for punctuation marks. The C7_Tagset is further upgraded to produce C8_Tagset to make finer distinctions in determiner and pronoun categories as well as for auxiliary verbs in English.

With regard to tagging guidelines, detailed strategies are proposed in CLAWS4 to decide how to draw the line of distinction between correct and incorrect assignment of tags. A guideline is essential as there are confusions with regard to the process of identifying the role of words in texts. A clearly defined guideline is required in tagging practice for the new set of scholars who are engaged in the task of text annotation. The guideline has been carefully created and instructed to remove confusion of any kind about what is a 'correct' or 'accurate' tag for a word in a corpus. Given below is a sample text tagged with CLAWS tagger (Figure 2).

Although the tagging guidelines proposed in CLAWS4 are useful for Indian languages, the tagset proposed in it needs to be customized to address the requirements of the Indian languages. The guidelines have few sets of language-independent rules which can be applied to any natural language. For instance, the major part-of-speech categories

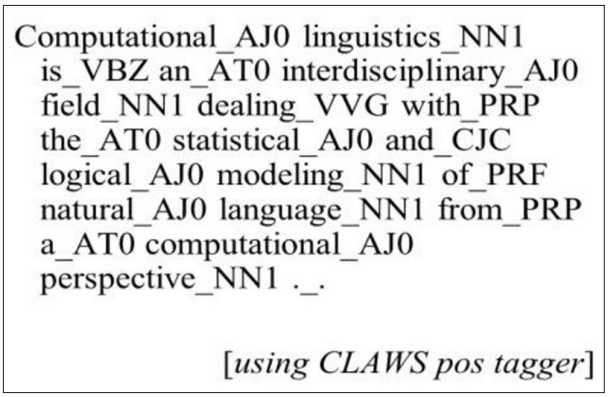

Figure 2. Screenshot of an output of CLAWS POS tagger (Courtesy: lancaster.ac.uk)

(e.g., *noun, verb, adjective, pronoun*) are more or less similar to all the natural languages. There is no need for modification of POS categories or the tagsets designed for these categories. However, modification is required for functional word categories (e.g., *indeclinable, conjunct, particle, determiner*) as well as for sub-categories of the main POS categories (e.g., *proper noun, common noun, abstract noun*). Necessary changes in the guidelines are needed to be incorporated in accordance with requirement of specific Indian languages. Only then we can think of using these guidelines for POS tagging in Indian language corpora.

HIDDEN MARKOV MODEL

In the middle of the 1980s, researchers in England, Norway, and Sweden, while working to tag a large database of the *Lancaster-Oslo-Bergen Corpus* (LOB), start using the Hidden Markov Model (HMM) to disambiguate parts-of-speech of words. The HMM is adopted for this purpose as it provides an opportunity for counting the cases (mostly from *Brown Corpus*) for developing tables regarding the probabilities of certain sequences of lexical items. Also, it provides vital cues in understanding the patterns of use of words in formation of natural sentences in texts. For instance, it helps researchers to trace that if the article *the* occurs at a certain place within a sentence, then the next word is either a noun (40%), or an adjective (40%), or a number (20%). This has been a piece of highly useful information for the annotators as well as for the system designers to decide about the patterns of the sequence of words normally occurring in a text. Based on information of this kind, system designer develops a program that can decide that the word *cook* within a string of *the cook has left for home* is far more likely to be a noun than a verb because its immediately preceding string is *the*. It helps annotators to benefit from the knowledge about the grammatical identity and part-of-speech of the words that follow immediately before or after a tagged word. Given below is a tentative architecture of the HMM model for POS tagging (Figure 3).

The higher-order HMM can do many more tasks by learning probabilities and possibilities. It not only helps to determine the possible word pairs but also provides cues to

know about those strings of three or more words that are to be combined together to form larger sequences. For instance, if the POS tagging algorithm encounters a determiner followed by a verb (e.g., $a_{\_DET}$ *barking*$_{\_VNF}$), then the possibility of the very next word to be a noun (i.e., $a_{\_DET}$ *barking*$_{\_VNF}$ *dog*$_{\_NN}$) is higher than being a verb, preposition or an article. Similarly, if there is a situation of one consecutive supporting verbs (i.e., *has been*), then the possibility of the very next word to be a verb (i.e., *has been working*) is higher than being a noun or an adjective. This approach is able to provide the much-required break-through in the development of automatic POS tagging programs for English and many other languages that follow similar syntactic rules (Kytö & Voutilainen 1995).

The application of the HMM approach to the *LOB Corpus* reveals some new findings. In certain situations where several ambiguous words occur together, the possibility of identification of actual part-of-speech of each word is multiplied. However, with higher-order HMM, it is possible to enumerate every acceptable combination of words. Also, it is possible to assign a relative probability to each combination by maximizing the probabilities of each acceptable combination. The combination that records the highest probability is chosen as the most suitable candidate for tag assignment. On an experimental basis, the team of *Lancaster University*, UK adopts this technique to achieve a considerably higher level of accuracy (93-95%) in the POS tag assignment on some non-customized English corpora. However, the main criticism against this approach is that in the act of dissolving ambiguities it merely assigns the most common tag to each known word which is against the standard practice of the POS tag assignment. For instance, the system assigns the tag 'proper noun' to all unknown words used in a text to achieve a good level of accuracy (nearly 90%). Many words are actually used within a text with other POS categories (Charniak 1997).

Another limitation of the HMM-based POS tagging system is noted for its 'generous nature' in the act of tag assignment. That means it goes for generating multiple possibilities of POS assignment to a single word or a single word combination without imposing any perceivable restriction. Although this technique generates a high rate of accuracy in the CLAWS system, it is expensive as it tends to enumerate all possibilities of multiword combinations. Moreover, sometimes, it resorts to backup methods when there are too many possibilities (Scott & Harper 1999). For example, the *Brown Corpus* contains an extreme example where all the words that occur within a string of seventeen words in a row are ambiguous. In such a situation, the HMM model generates possible POS combinations in the order of thousands. It becomes a challenging task both for a human annotator and a system to verify all the combinations and approve the acceptable ones. Also, there are cases where the word *still* is represented in as many as twelve different parts-of-speech. Such complexities in identification of ambiguities by HMM program make a POS tagged corpus more deceptive and less reliable than it actually is before it is put to tagging (Pajarskaite 2004).

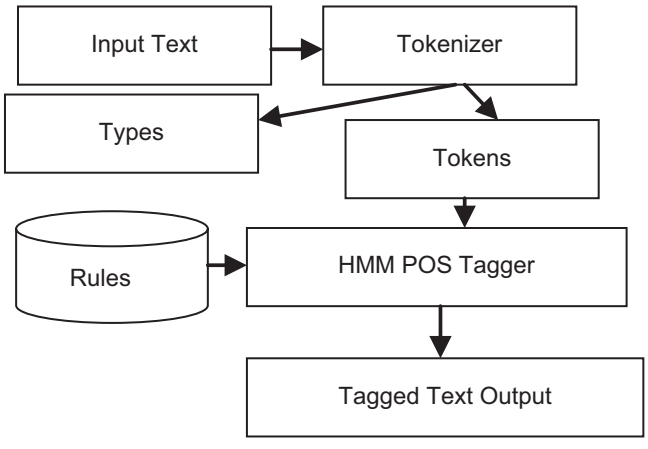

Figure 3. HMM based POS tagging architecture

DYNAMIC PROGRAMMING ALGORITHM

A comparatively new method, namely, Dynamic Programming Algorithm (DPA) is developed by scholars to dissolve problems of ambiguity in the POS tag assignment to words in corpora (DeRose 1988). The main advantage of this method is that the algorithm can execute the task of tag assignment within a short span of time and with a better accuracy rate. Based on the Viterbi Algorithm (Forney 1973) this method uses a table of word pairs in an indigenous way to estimate values for word-pairs in a corpus. Following this strategy, it achieves not only a high rate of accuracy (over 95%) in a trial corpus, but also includes within its analysis the results of specific types of error, probabilities, and other related information (Abney 1997). The strategy is adopted for the English text corpus is also replicated on a Greek language corpus. In this case also, it produces much better results than the previous methods. Given below is a tentative diagram (Figure 4) of the DPA to show how the system operates in POS tagging.

This innovative POS tagging method disrupts many on-going POS tagging activities in English and other languages. The rate of accuracy reported in this study (DeRose 1990) is much higher than the typical level of accuracy that is acquired through application of existing algorithms that usually integrated part-of-speech information of words with other levels of linguistic information relating to syntax, morphology, semantics, and so on (Britto *et al.* 1999).

Although methods like HMM are accepted as standard methods for POS tagging, many new methods are also

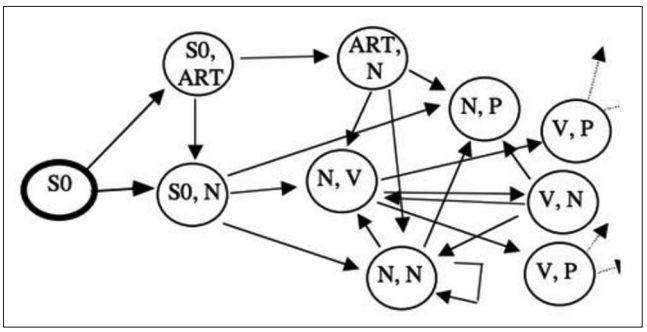

Figure 4. Dynamic programming algorithm based on the viterbi algorithm

considered with the introduction of rule-based, stochastic, neural, statistical, Support Vector Machine (SVM), Maximum Entropy Classifier, Perceptron, Nearest Neighbour, and deep learning-based approaches. Most of the new approaches like Brill Tagger and TnT Tagger promise to achieve a higher level of accuracy. Two of them (Brill Tagger and TnT Tagger) are briefly discussed in the following two sections.

BRILL POS TAGGER

The Brill POS Tagger is developed by Eric Brill in 1993. It is also known as an 'error-driven transformation-based tagger', which generates an interface in the act of doing part-of-speech tagging in a pre-defined rule-based method (Brill 1992). It is error-driven in the sense that it recourses the process of supervised learning, and it is transformation-based in the sense that a tag may be assigned to a word and the tag may be changed using a set of pre-defined rules. For instance, if a word is already known to the system, the tagger first assigns the most frequent tag. On the other hand, if the word is unknown to the system, the tagger naively assigns the tag of 'noun' to it. Thus applying the rules over and over and changing the incorrect tag when required, this system is able to achieve a high level of accuracy. The algorithm used by this method can be summarised in the following six stages:

Stage 1: Start the POS tagger on a digital corpus database.
Stage 2: Encounter a (new) word (in the inbuilt lexicon) and assign the most frequent tag associated with the word.
Stage 3: Encounter an unknown word (out of inbuilt lexicon) and tag it as a proper noun if capitalized, else as a simple noun (if not capitalized).
Stage 4: Learn or guess tags on the basis of contextual rules.
Stage 5: Change the incorrect tag to a correct one with contextual rules.
Stage 6: Generate output.

The initial learning phase of Brill POS Tagger involves several sub-stages and strategies as the following. A tentative diagram of the method is given hereafter (Figure 5):

(a) First, it iteratively computes the error score of each candidate rule to calculate the difference between the number of errors before and after applying the rule.

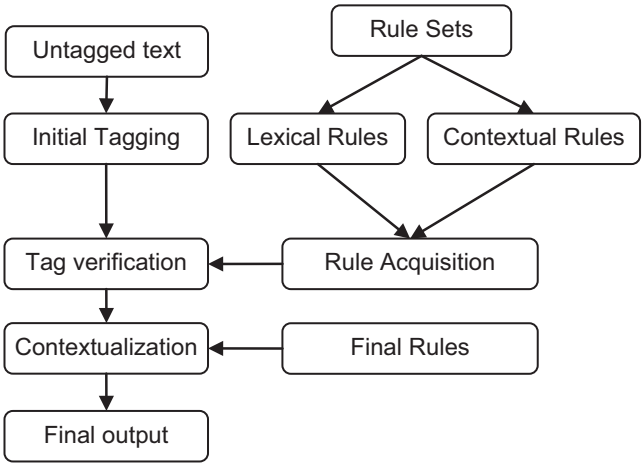

Figure 5. Simplified architecture of brill POS tagger

(b) Second, it selects the best (higher score) rule.
(c) Third, it adds it to the rule set and applies it to the text again.
(d) Fourth, it repeats the process until no rule has a score above a given threshold. That means, if the chosen threshold is zero, it continues the application of rules until it achieves a greater score than the chosen threshold. Once it is achieved, it is then supposed to be the final stage of the tagging.

For achieving a greater level of success it applies two sets of rules: the first set of rules is called 'Lexical Rule', which is used for initialization of the process. The second set of rules is known as 'Contextual Rule', which is used to remove errors and correct the final tags.

(a) Lexical Rule : Tag W_1 → W_N IF W_1 carries suffix like '-tion', '-ment', etc.
(b) Contextual Rule : Tag W_1 → W_2 IF the preceding or the following tag is X.

The Brill POS Tagger, is, however, not problem-free. It has problems with parts-of-speech of those words that belong to open class. Words of this kind have a complex lexicosyntactic identity due to which they are free to perform roles of different parts-of-speech in different contexts. On the other hand, based on existing standard grammatical categories, words of a closed class are easier to detect and annotate correctly. It is noted that it is more difficult to derive information from a word which is annotated with a simple tag denoting its primary part-of-speech, than from a word whose tag includes information about its morphological structure and inflectional properties. Therefore, it is argued that for better evaluation of a POS tagger it is necessary to test the system on several corpora where words are used with varied morphosyntactic and grammatical information. High accuracy of a system is achieved by applying it to larger corpora made with texts taken from different subject domains (Nissim *et al.* 2004). On the other hand, a reference to a large lexical database can reduce the number of unknown words, which can be a good advantage for the system. Furthermore, the level of accuracy of a system can be greatly improved if the rule generating mechanism is flexible enough to consider different morphological and lexicosyntactic characteristics of words of a language.

Although it is argued that Brill POS Tagger is a language-independent system (Brill 1995), in actuality, it faces difficulties with those languages which are characteristically different from English. The words used in Indian languages, for instance, are dissimilar in form and characteristics from that of English. We argue that if Brill POS Tagger is adopted for the Indian languages, it will be a tough challenge for the system to perform as it will face hurdles with regard to orthographic representation, morphological structure, and lexicosyntactic functions of words.

TNT POS TAGGER

*The Trigrams-n-Tagger (TnT) is claimed to be an elegant and e*fficient system of statistical POS tagging that can be trained for different languages and can virtually be adopted for any tagset usable for a language (Brants 2000). The

level of performance of this system depends largely on its ability in generating components for parameters that are developed through its use on previously tagged corpora. It is also claimed that the system is capable of incorporating several methods of smoothing to handle unknown words that are not previously encountered in texts. Moreover, since the tagger is not optimized for a particular language or a variety, it can be trained with data taken from a wide variety of texts belonging to different languages. Therefore, there are fewer problems for this system to adopt a new tagset, to deal with a new genre of a text, and to apply on texts of a new language. The algorithm can be optimized for speed so that it can generate quick outputs from all kinds of text on which the tagger is applied.

The TnT tagger is a combination of the Viterbi Algorithm for the second-order Hidden Markov Model. It processes words by analyzing their suffix parts that are attached to words. Primarily, looking at the specific trailing part which is tagged to a word, it tries to determine the part of speech of a word. The primary paradigm that is used for smoothing is a linear interpolation while respective weights are determined by delayed interpolation on the information derived from linear interpolation. In this system, therefore, all unknown words are tagged based on their suffix part and successive abstraction. In case of those words where the suffix part is missing, it depends on delayed interpolation empowered with information derived from lexical databases. The tagger may be applied to a text by using the following three modes:

(a) **Initial Mode:** Input file will contain one token per line.
(b) **Base Mode:** Tagger adds the second column to each line, containing the tag for the word.
(c) **Third mode:** Tagger emits alternative tags for each token, together with a probability distribution.

In the output database, if a word is marked with an asterisk (*), it has to be considered that the word is not in the lexicon used by the TnT tagger. In that case the database has to be augmented with a new lexical dataset as well as the dataset has to be populated with additional information relating to their possible part-of-speech affiliation. The speed of the tag assignment to words in text largely depends on the amount of ambiguity of words and the percentage of unknown words used in a text. Given below is a format of the untagged and tagged file in TnT tagger (Figure 6).

This tagger is applied on a small part of the *Susanne Corpus* and it generates 94.5% accuracy due to small size of a corpus (around 1,50,000 tokens) and large tagset (around 160 plus multi-token tags). When it is applied to large English corpora like *Penn Treebank*, the accuracy of the output is much higher (96.7%) as the number of tagset is reduced. It is claimed that the tagger can be trained on different language databases where written words are separated by white space (Brants 2000). A notable limitation of the TnT tagger is that it acts well with any tagset represented in ASCII, but cannot work properly where tagset is represented in Unicode (UTF-8).

%% Brown Corpus	%% Brown Corpus
%% File N11, Sent 3	%% File N11, Sent 3
But	But CC
the	the DT
day	day NN
of	of IN
the	the DT
deadline	deadline NN
came	came VBD
and	and CC
passed	passed VBD
,	, ,
and	and CC
the	the DT
men	men NNS
who	who WP
had	had VBD
scoffed	scoffed VBN
at	at IN
the	the DT
warnings	warnings NNS
laughed	laughed VBD
with	with IN
satisfaction	satisfaction NN
.	. .
a) untagged format (one column)	*b) tagged format (two columns, separated by white space)*

Figure 6. Format of untagged and tagged files in TnT tagger

FINDINGS AND DISCUSSION

Almost all the POS tagging methods and approaches discussed above use pre-existing digital language corpora for trial and experiment as well as for verification of the methods that are applied to actual texts. Many of the experiments reveal that it is very much possible to bootstrap language texts by using 'unsupervised' tagging conventions particularly when the amount of natural text data is very big in size and varied in content. In most cases, an unsupervised tagging technique uses a small part of the untagged corpus for training purposes and produces the tagset through induction. The system observes the patterns of word use in a sample database and derives part-of-speech values for words through cross-reference. For example, statistical information reveals that English particles *the, a,* and *an* can occur in similar syntactic contexts, verb *eat* can occur in a context that is different from that of particles. The application of this technique with sufficient iteration and repeated use in various texts can generate similarity classes of words that are similar to those what humans expect or design for. Also, the application of this technique sometimes produces certain differences which provide valuable insights about the parts-of-speech of words which can never be presumed by human annotators.

For decades now, POS tagging is treated as an inseparable part of text processing. It is an indispensable strategy for any system or tool that is used in language technology to identify accurately part-of-speech of words in a piece of text. To date, we have come across a few POS tagging systems, which are quite dynamic, workable and robust. This is true for many of the advanced languages like English, Spanish, and German. However, our regret is that there is not a single POS tagger which is full-proof, automatic, robust, and ambiguity-free for resource-poor languages like Bengali, Tamil, Hindi or Marathi.

What is understood from the discussion presented above is that there are large numbers of words in natural texts for which assignment of correct part-of-speech is difficult. The system cannot assign appropriate part-of-speech value blindly without understanding the contextual occurrence of words and meaning of these words denoted in different contextual frames (Smith 1997). We also need to understand the pragmatic and discoursal roles of words in texts for assigning the right POS tag to them.

Part-of-speech tagging is a complex and expensive process. Pre-processing of texts is a necessary pre-condition for a tagging system as it can generate information and knowledge that are required for analyzing deeper levels of text analysis relating to semantics, syntax, pragmatics, and discourse. Therefore, developing a workable POS tagger that can run on texts and generate correct outputs is a dream that is yet to be realized for many of less advanced languages across the world.

In general, however, all the methods and algorithms that have been designed and developed so far have largely failed in case of those words where information from the domains like semantics and pragmatics is required for tag assignment. The basic challenge is that a word, when put to a context, automatically absorbs much of the information from its context of use due to which its actual contextual part-of-speech tends to change from the part-of-speech that is recorded in structured lexical resources like dictionaries and grammars. Strikingly, such cases are not rare in a natural language as words in every language carry this typical linguistic feature. Therefore, most of the systems fail in those cases where information from domains of discourse, pragmatics, and extralinguistics is required for identification of part-of-speech of words.

Such limitations in the POS tag assignment led us to realize that part-of-speech tagging should strictly be separated out from other levels of corpus annotation, such as syntactic structure annotation (i.e. parsing). It is also understood that part-of-speech annotation and syntactic structure annotation are two different ways of treating a natural language text. Besides, they have different goals and different operational methodology, and hence, they should be treated separately in corpus processing. This strategy can simplify our approaches towards looking at the texts of a language as well as encouraged researchers to separate part-of-tagging method from other methods of corpus processing to chalk out a new way for this particular task.

VALUE OF A POS TAGGED CORPUS

The importance of a POS tagged corpus is enormous in language description, language description, language computation and language cognition. The applicational role of a POS tagged corpus, within a wider canvas of descriptive, computational, cognitive, and applied linguistics is visualized in the following diagram (Figure 7). It shows how a POS-tagged corpus is a primary source of data and information for linguistic works of various kinds.

POS tagging is the first step for most of the complex language processing applications like developing systems for grammar checking, named entity recognition and extraction, word sense disambiguation, sentence parsing, text understanding, query addressing, information retrieval, machine translation. For major works of language processing, a tagged text corpus is useful for extracting linguistic data and grammatical elements to be used in machine learning and language modeling. In the area of descriptive and applied linguistics, a POS tagged text is useful for works of frequency calculation, type-token analysis, lemmatization, lexical

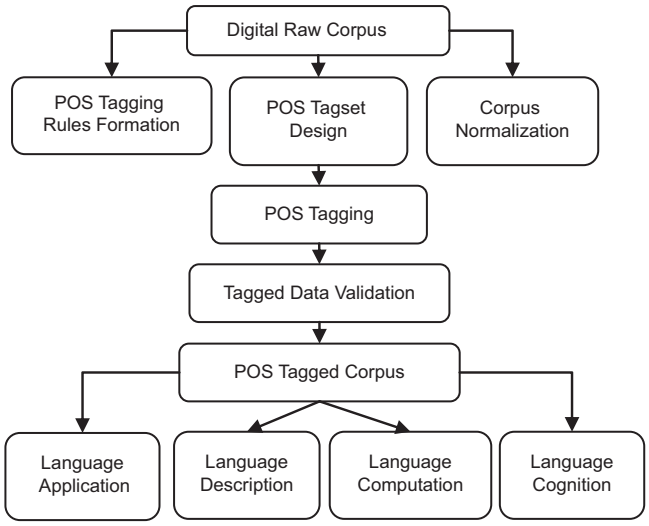

Figure 7. Tentative architecture of POS tagged corpus generation

sorting, vocabulary list compilation, dictionary-making, and language teaching.

We can visualize many applications of a POS tagged corpus in a language. We can also expect that we can take trouble to develop this system for Indian languages as it has high functional relevance in these languages. In the recent past, efforts are made to develop a POS tagger for those Indian languages that are included in the ILCI (Indian Languages Corpora Initiative) project for developing parallel translation corpora in Indian languages. However, rather than developing an unsupervised system, we have developed a tool that can be used by human experts to tag words in texts with a high rate of accuracy.

On the other hand, some of the automated POS tagging systems for the Indian languages cut a sorry figure. For instance, in a corpus of a hundred thousand Bengali words, the rate of accuracy is 85 to 90% (Dandapat 2009). In contrast, in a one-million word English text of the *American National Corpus* the rate of accuracy is over 98%. This indicates that we need to take serious initiative in this direction to develop POS tagged system for the Indian languages with two specific goals: (a) we have to design a large tagset to increase the rate of accuracy of a POS tagging system, and (b) we have to customize the system taking into consideration the unique linguistic and morphological properties of words of a language so that the system is language-specific and accurate in generation of outputs.

ACKNOWLEDGMENT

The authors extend their sincere thanks and appreciation to the Deanship of Scientific Research and Research Centre, College of Arts, King Saud University.

REFERENCES

Abney, S. 1997. Part-of-speech tagging and partial parsing. In: Schreibman, S., Siemens, R.G. & Unsworth, J.M. eds. *Corpus-Based Methods in Language and Speech:*

A Companion to Digital Humanities. London: Blackwell. Pp. 118-136.

Archer, D. & Culpeper, J. 2003. Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics. In: Wilson, A., Rayson, P. & McEnery, A.M. eds. *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Peter Lang: Frankfurt. Pp. 37-58.

Archer, D., McEnery, T., Rayson, P., & Hardie, A. 2003. Developing an automated semantic analysis system for Early Modern English. *Proceedings of the Corpus Linguistics 2003 conference*. UCREL, Lancaster University. Pp. 22-31.

Archer, D., Rayson, P., Piao, S., & McEnery, T. 2004. Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In: Williams, G. & Vessier, S. 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 III. Pp. 817-827.

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

Biber, D., Finegan, E., & Atkinson, D. 1994. ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers. In: Fries, U., Tottie, G. & Schneider, P. eds. *Creating and Using English Language Corpora*. Amsterdam: Rodopi. Pp. 1-14.

Brants, T. 2000. TnT- A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000), Seattle,* WA, USA. Pp. 37-42.

Brill, E. 1992. A simple rule-based part of speech tagger. *Proceedings of the Workshop on Speech and Natural Language (HLT-91)*, Morristown, NJ, USA: Association for Computational Linguistics. Pp. 112-116.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics.* 21(4): 543-565.

Britto, H., Galves, C., Ribeiro, I., Augusto, M., & Scher, A. 1999. Morphological Annotation System for Automatic Tagging of Electronic Textual Corpora: from English to Romance Languages. *Proceedings of the 6th International Symposium of Social Communication*. Santiago, Cuba. Pp.582-589.

Charniak, E. 1997. Statistical Techniques for Natural Language Parsing. *Artificial Intelligence Magazine*. 18(4): 33-44.

Dandapat, S. 2009. *Part-of-Speech tagging for Bengali*. MS Thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India (MS). Dash 2011).

Dash, N.S. 2011. Principles of Part-Of-Speech (POS) Tagging in Indian Language Corpora. In: Vetulani, Z. ed. *Proceedings of 5th Language Technology Conference (LTC-2011): Human Language Technologies as a challenge for computer science and linguistics.* Poznan, Poland, 25-27 November 2011, Pp. 101-105.

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

DeRose, S.J. 1990. *Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages*. Doctoral Dissertation, Department of Cognitive and Linguistic Sciences, Providence, RI: Brown University, USA.

Fligelstone, S., Pacey, M. & Rayson, P. 1997. How to generalise the task of annotation. In: Garside, R., Leech, G. & McEnery, A. Eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman. Pp 122-136.

Fligelstone, S., Rayson, P. & Smith, N. 1996. Template analysis: bridging the gap between grammar and the lexicon. In: Thomas, J. & Short, M. eds. *Using Corpora for Language Research.* Harlow: Longman. Pp 181-207.

Forney, G.D. 1973 The Viterbi algorithm. *Proceedings of the IEEE*. 61(3): 268-278. doi:10.1109/PROC.1973.9030.

Garside, R. & Smith, N. 1997. A hybrid grammatical tagger: CLAWS4. In: Garside, R., Leech, G. & McEnery, A. eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman. Pp. 102-121.

Garside, R. 1987. The CLAWS Word-tagging System. In: R. Garside, Leech, G. & Sampson, G. eds. *Computational Analysis of English: A Corpus-based Approach*, London: Longman. Pp. 30-41.

Garside, R. 1995. Grammatical tagging of the spoken part of the British National Corpus: a progress report. In: Leech, G., Myers, G. & Thomas, J. eds. *Spoken English on Computer: Transcription, Markup, and Application.* London: Longman. Pp. 161-167.

Garside, R. 1996. The robust tagging of unrestricted text: the BNC experience. In: Thomas, J. & Short, M. eds. *Using corpora for language research: Studies in the Honour of Geoffrey Leech*, London: Longman. Pp. 167-180.

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

Kytö, M. & Rissanen, M. 1993. General introduction. In: Rissanen, M., Kytö, M., & Palander-Collin, M. eds. *Early English in the computer age: explorations through the Helsinki Corpus*. Berlin: Mouton de Gruyter. Pp. 1-17.

Kytö, M. & Voutilainen, A. 1995. Applying the Constraint Grammar Parser of English to the Helsinki Corpus. *ICAME Journal* 19: 23-48.

Leech, G. & Eyes, E. 1993. Syntactic annotation: linguistic aspects of grammatical tagging and skeleton parsing. In: E. Black, Garside, R. & Leech, G. eds. *Statistically-driven Computer Grammars of English: the IBM/ Lancaster Approach*. Amsterdam: Rodopi. Pp. 36-61.

Leech, G. 1997. Introducing Corpus Annotation. In: Garside, R., Leech, G., & McEnery, A. eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman. Pp. 1-18.

Leech, G., Garside, R. & Atwell, E. 1983. The automatic tagging of the LOB Corpus. *International Computer Archive of Modern English News*. 7(1): 110-117.

Leech, G., Garside, R. & Bryant, M. 1994. CLAWS4: The tagging of the British National Corpus. *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)* Kyoto, Japan. Pp. 622-628.

Leech, G., Rayson, P., & Wilson, A. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus.* Longman, London.

Mueller, M. 2005. The Nameless Shakespeare. *Working Papers from the First and Second Canadian Symposium on Text Analysis Research (CaSTA)*. Computing in the Humanities Working Papers (CHWP 34).

Nissim, M., Matheson, C. & Reid, J. 2004. Recognizing Geographical Entities in Scottish Historical Documents. *Proceedings of the Workshop on Geographic Information* Retrieval at SIGIR 2004.

Osselton, N.E. 1984. Informal spelling systems in Early Modern English: 1500- 1800. In: Blake, N.F. & Jones, C. eds. *English Historical Linguistics: Studies in development.* The Centre for English Cultural Tradition and Language, University of Sheffield, Pp. 123-137.

Pajarskaite, G. 2004. Designing HMM-based Part-of-Speech Tagging for Lithuanian Language. *Informatica*. 15(2): 231-242.

Piao, S.L., Rayson, P., Archer, D., & McEnery, T. 2004. Evaluating lexical resources for a semantic tagger. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004),* 26-28 May 2004, Lisbon, Portugal, Volume II, Pp. 499-502.

Rayson, P., Archer, D., & Smith, N. 2005. VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora. *Proceedings of Corpus Linguistics 2005, Birmingham University,* July 14-17, 2005.

Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. *Technical Report MS-CIS-90-47,* Department of Computer and Information Science, University of Pennsylvania.

Scott M.T. & Harper, M.P. 1999. A second-order Hidden Markov Model for part-of-speech tagging. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* Pp: 175-182.

Siemund, R. & Claridge, C. 1997. The Lampeter Corpus of Early Modern English Tracts. *ICAME Journal*, 21: 61-70.

Sinclair, J. 2004. *Trust the Text: Language, Corpus, and Discourse.* London and New York: Routledge.

Smith, N. 1997. Improving a tagger. In: Garside, R., Leech, G. & McEnery, A. eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman. Pp. 137-150.

WEB LINKS

http://ccl.pku.edu.cn/doubtfire/NLP/Lexical_Analysis/Word_Segmentation_Tagging/CLAWS/CLAWS%20part-of-speech%20tagger.htm

http://drops.dagstuhl.de/opus/volltexte/2007/1055

http://ucrel.lancs.ac.uk/claws5tags.html

http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2postag_manual.htm

https://dictionary.cambridge.org/dictionary/english/still

https://www.researchgate.net/publication/2618590_The_CLAWS_Web_Tagger

https://www.sketchengine.eu/english-claws5-part-of-speech-tagset

http://www.coli.uni-saarland.de/~thorsten/publications/Brants-TR-TnT.pdf