4.0 INTRODUCTION

The results of a morphological study are essential from the point of almost all practical applications that deal with natural language. After all, an application must first recognize the word in question before analyzing it syntactically, semantically, or whatever the level may be (Anton 2001). In this chapter, we describe the importance of the statistical study of morphology and how it can be used in various Natural Language Processing applications for Tamil. The morphological strength of Languages demands the use of thorough morphological analysis. Morphological analysis should be the first step towards any language processing task. We discuss here how this morphological analysis of the language assists in building and testing of NLP applications and enhancing their coverage. The results of the statistical analysis of Tamil morphology would be useful for enhancing the NLP applications such as WSD, Morphological Analyzer, Generator, POS tagger, Spell Checker, Finite State Automaton and MT etc.

The process of segmenting words into morphemes involves the study of word formation. The Tamil language is morphologically rich and agglutinative too. Each root word is affixed with several morphemes to generate word forms. Tamil involves rightward inflection to the root word. Computationally each root word can take hundreds of suffixes and generate a great number of inflected word forms, out of which only few thousands or hundreds exist in any typical corpus. For the purpose of analysis of such inflectionally rich language, the root and the morphemes of each word have to be studied from the point of its distributional aspects.
4. 1 MORPHOLOGICAL ANALYSIS OF TAMIL

The morphological structure of Tamil is quite unique and complex since it inflects for person, gender, and number markings and also combines with auxiliaries that indicate aspect, mood, causation, attitude etc., in case of verb and number, case, postpositions in case of nouns. For the purpose of analysis of such inflectional formations, the root and the morphemes of each word have to be identified based on corpus. The structure of verbal complex is unique and understanding of this will be a great helpful for any researcher as well as for a machine. The formation of the verbal complex involves arrangement of the verbal units and the interpretation of their combinatory meanings. Phonology also plays its part in the formation of verbal complex in terms of morphophonemic or sandhi rules which account for the shape changes due to inflection and classification of Tamil verbs is based on tense inflections. The inflection includes finite and non-finite and the later involves infinitive, adjectival, adverbial and conditional forms of verbs. Compared to verbal morphological analysis, noun morphological analysis is less challenging. Nouns occur with plural inflection, oblique formation, case, postpositions and clitics. (Rajendran et al. 2001).

Natural Language Processing is one of the most rapid developing areas nowadays. For a majority of languages, pools of raw texts or even simple wordlists are the resources most urgently needed. Statistical studies of corpus data are, for many applications, the raw fuel of NLP, and the testbed on which an NLP application is evaluated. It is the main requirement for almost every kind of natural language processing, such as machine translation etc. So, since 1960, when one of the first instances of a text corpus was created (Brown corpus), many such corpora were designed, together with the methods that were used in their production. In the last decade, new approaches in designing textual corpora have been developed of which crawlers of the text from the web. The later comes near the idea of having a
heterogeneous and informal language of communication. And this also serves another purpose - it democratizes the way the linguists work.

Corpus-based statistical studies to language have introduced new dimensions to linguistic description and to various applications by permitting some degree of automatic analysis of text. The most basic format used in displaying information about the linguistics elements in a corpus is generated by means of listing and counting.

One of the most explored fields of NLP is morphology. It is important because language is productive. In any given, one would encounter text that has a number of words and word forms that we haven’t seen before and that are not in any precompiled dictionary. The core task of computational morphology is to take word as input and produce a morphological analysis for it. Morphotactics defines concatenation of the principal morphemes of a word, and it is typically described through finite state automata. But there are situations where the word formation process is not just joining of morphemes, besides other processes like reduplications, incorporation etc. These will be situations where the phonological rules are called for ensuring well-formedness. Phonological rules may apply and change the shape of morphs. Many linguists have modeled phonological rules, but it is considered that the most successful one is the model called two-level morphology (Koskenniemi, 1983). The two-level morphology model has been proved successful for formalizing the morphology of very different languages (English, German, French, etc.). This system is used even for conversion between different writing systems, it seems too.

Natural languages consist of a very large number of words that are built upon basic building blocks known as morphemes, the smallest linguistic units to possess meaning. In natural language processing, Morphology concerns with the discovery and analysis of the internal structure of words using computers. The result of the statistical study of morphology is relevant
in the selection and application in Word Sense Disambiguation, Morphological Analyzer, Morphological Generator, Parts-Of-Speech, Machine Translation, Spell-checker, Finite State Automaton etc.

Users of natural languages need no persuasion that words are naturally occurring units. We may have a doubt as to whether expressions like the māṭṭu vaṇṭi in Tamil and bull-cart in English should be treated as one word or two, but there is no disagreement about the notion that sentences can be analytically broken down into component words. Linguists and others who think deeply about such questions as “what is a word?” There are some difficult intermediate cases which linguists call clitics, particles, morphemes whose status as a full-fledged word is dubious. In all languages, or virtually all, it is appropriate to analytically break words down into component pieces, called morphs, and then to bundle morphs back into the functional units we call morphemes. Such an analysis is part of the functionality of a morphology, and is the central subject of this chapter (when a morpheme corresponds to only a single morph, as is often the case, we generally ignore the difference between a morph and a morpheme). In addition, an exhaustive description of Tamil morphology would be necessary to identify an appropriate set of morphosyntactic features of a word; to the extent that the word’s morphological decomposition can serve as a basis of specifying those features. Thus in this work, words are in terms of affixes, and the affixes should be labeled as indicating prefix + root + suffixes; plurality, case marker, clitics for nouns and tense, aspects, mood number, gender for verbs.

Implementing sophisticated computational accounts of natural language morphology goes back more than forty years. There have been several excellent book-length studies of computational morphology in recent years, with considerable concern for actual, real-world implementation, notably by Beesley & Karttunen (2003), Uma Maheshwar Rao (2006) and Roark & Sproat (2007), the grammar of words by Geert Booij (2007), as well as Ritchie et al. (1992) and Sproat (1992). Current work in this area focuses to a
USE OF MORPHOLOGICAL STUDY IN NLP APPLICATIONS

very large extent on natural language processing and the use of finite-state transducers as a means to carry out the functions of morphology. This work was stimulated by the work of Douglas Johnson (1972), Ronald Kaplan and Martin Kay (1981), Kimmo Koskeniemi (1983) and Lauri Karttunen and colleagues (1993). A well studied morphology of a language can be vital for many practical applications such as WSD, MA, MG, POS, MT, FSA, Spell-checker etc. A very humble but honest function of many products was appreciated by a wide range of end-users, and in a language with a rich inflectional morphology, as we know in languages such as Telugu, Tamil, Finnish, Hungarian and many others. The total number of possible forms that a user might reasonably generate is far greater than the capacity of a computer to hold in its memory, search and retrieve unless the entire family of forms is compressed to a manageable size by virtue of the redundancies inherent in a computational morphology. It is typically the inflectional morphology which gives rise to the very large number of possible forms for nouns and verbs, and it is typically the inflectional morphology which can most usefully be stripped off when one wishes to build a NLP tool based not on actual words, but on the most useful part of the words. Syntactic parsing in most languages requires knowledge of the morphosyntactic features carried by each word, and that knowledge is generally understood as being wrapped up in the morphology (primarily the inflectional morphology) of the language. A number of researchers have explored the effect of statistical study of morphology for the NLP applications that are produced by incorporating knowledge of inflectional and derivational morphology (Harman (1991), Krovetz (2000), Hull (1996), Kraaij & Pohlmann (1996), Xu & Croft (1998), Goldsmith et al. (2001), Larkey (2002), and Savoy (2006)).

4.2 APPLICATIONS

The results of the statistical study and their relevance in the selection and application in NLP are described in the following sections.
4.2.1 POS TAGGING

A word can be classified into one or more lexical or parts-of-speech categories such as nouns, verbs, adjectives, and adverbs, to name a few. Parts of speech (POS) tagging is the process of labeling annotation of syntactic categories for each word in a corpus. In other words, POS tagging is the process of labeling a part of speech or other lexical category to each and every word in a sentence. Tagging is also known as the primary phase in the assignment of structure to a text. For example, *avan intap peñai nēcikkirān.*

Once the data is tagged either at a morphosyntactic or POS level then the State-of-the-art POS taggers can achieve a good accuracy using statistics to determine its function. Tagging text with parts-of-speech turns out to be extremely useful for linguistic research and more complicated NLP tasks such as parsing and machine translation are developed on this.

For some time, parts-of-speech tagging was considered as an inseparable part of natural language processing. Developing a correct parts-of-speech without understanding the semantics or even the pragmatics of the context could be extremely difficult. This is extremely expensive, especially because analyzing the higher levels is much harder when multiple parts-of-speech possibilities ought to be considered for each word.

In this context, it is impossible to think over research on parts-of-speech tagging without corpus. Corpus has been used for innumerable studies involving parts-of-speech, and also has inspired the development of similar "tagged" corpora in many other languages. However, now it has been superseded by involving larger corpora such as the 100 million words British National Corpus. For a corpus, there is a necessity for POS tagging so that the resources can be utilized in natural language parsing, machine translation, etc.
4. 2. 2 MORPHOLOGICAL ANALYZER

Morphological analysis is the process of segmenting words into morphemes or analyzing the process of word formation. It is a primary step for various types of text analysis of any language. Morphological analyzer takes a word as an input and produces the root, and its grammatical features as the output (Uma Maheshwar Rao, 2006). For example,

Input: oxen (English)

Output: [root = ox, category = n, number = plural]

It is used as one of the components in machine translation etc. Corpus is used in the development of morphological analyzer for enhancing its coverage. The partially developed tagged corpus can be used to improve morphological analyzer using bootstrapping methods.

Morphological analysis is a very significant step towards efficient NLP for highly inflectional languages like Tamil. Morphology is one of the complementary parts of the structural aspects of natural language expression. This study is particularly significant because, apart from acquisition of morphology we could focus on the natural language generation and also on the possible decompositions of the given word with probability of that occurrence. The Unsupervised Morphological analyzer can be improved by making it semi supervised. That is adding the learning algorithms which could learn the sandhi rules based on the output generated by the Rule based analyzer and applying these rules while decomposing. If such strategy is used then with small data we can generate large vocabulary and also improve the performance of a morphological analyzer.

4. 2. 3 MACHINE TRANSLATION

As we mentioned earlier that morphological study is an important task shared by many applications in the field of NLP. Closely connected to this
task, one of the main issues at hand when designing such applications is how to organize and store in the lexicon the morphological information needed to analyze and generate words. Machine Translation is typically one of the applications that need to handle analysis and generation processes at the same time. Linguistic databases for MT systems need to be designed so that the knowledge they store is as much process independent as possible. Thus designing declarative lexicon databases in MT applications is a must.

Machine Translation (MT) is a task with multiple components, each of which can be very challenging. MT makes it possible and easier to collect knowledge in other languages, as well as to distribute knowledge to other languages. Parallel corpus has been used as an aligned pair of any two languages in MT. The aligned data comes from a corpus. Parallel corpora involves two (or more) corpora in different languages, each containing text that have been translated from one language into to another (e.g. a novel in English that has been translated into Tamil, and one in Tamil that has been translated into English) or texts that have been produced simultaneously in two or more languages. They can be used by translators and by learners to find potentially equivalent expressions in each language and to investigate differences between languages. Some parallel corpora include European Union regulations, which are published in all the official languages of the European Union.

Using corpus approach to machine translation usually begins with a bilingual training corpus. This approach is to extract from the corpus for generalized statistical knowledge that can be applied to new, unseen test sentences. A different approach is to simply memorize the bilingual corpus. This is called translation memory, and it provides excellent translation quality in the case of a "hid" (i.e., a test sentence to be translated has actually been pre-observed in the memorized corpus).
MT systems may use a large monolingual corpus to improve the accuracy of translated words, phrases/sentences. The MT system may produce alternative translations and use the large monolingual corpus to (re)rank the alternative translations. The MT system may receive an input text segment in a source language, compare alternate translation for the said input text string in a target language and record a number of occurrences of the alternate translations in the large monolingual corpus. The MT system may then re-rank the alternate translations based, at least in part, on the number of occurrences of each translation in the corpus.

Statistical MT

What linguistic knowledge sources are needed to perform MT depends at least to a limited extent, on the method that is used. For example, some US projects (Brown et al., 1989) made use of very large scale statistical information from texts, while Japanese systems do not. Conversely, an experimental MT system at Kyoto University made use of large lists of sample sentences against which a sentence to be translated is matched (Nagao, 1990), whereas no US systems did this until the turn of the century (the CMU system). Most MT systems however, make use of at least some, or possibly all of the following kinds of lexical knowledge sources as distinct from corpora alone (as in Brown et al.). The morphological analysis has been shown to improve the results in automatic translation between languages with different morphological structures IN Indian Languages (Uma Maheshwar Rao and Amba K., 2008).

4. 2. 4 WORD SENSE DISAMBIGUATION

Word sense disambiguation is defined as the task of finding the correct sense of a word in a context. This is crucial for applications like Machine Translation and Information Extraction. Sense-tagged corpus is still not large enough to create the building of a wide coverage, high accuracy Word Sense
Disambiguation (WSD) program that can significantly outperform the most-frequent-sense classifier over all content words encountered in an arbitrarily chosen unrestricted text. The amount of human annotation effort needed can be considered as an upper bound on the manual effort needed to construct the necessary sense-tagged corpus to achieve wide coverage WSD using statistics. It may turn out that we can achieve our goal with much less annotation effort.

Large text corpora and the computational resources to handle them have recently become available to computational linguists. In order to apply to multi-million word corpora, natural language processing techniques must be efficient and domain-independent; for this reason, coarse or partial analyses are becoming more attractive. For example, coarse syntactic interpretation, such as partial parsing (de Marcken 1990) (McDonald 1990) and automatic collocation generation (Smadja & McKeown 1990) (Choueka 1988) are being explored.

**Bootstrapping From Bilingual Corpora**

The time required for hand-labeling the training sentences are prohibitive, but there is a way it might be automated. Recently several researchers (e.g. (Brown et al. 1991), (Dagan et al. 1991)) have suggested using bilingual aligned corpora in the lexical disambiguation task, (the term “aligned” indicates that within the bilingual database, sentences that are translations of one another are grouped together). Although most of the discussion is in terms of choosing words for translating one language to another, it is also suggested that, because the words that have more than one sense in language A have only one sense or a different set of senses in language B, by using a most frequently occurring bilingual dictionary, the correct sense of a word in language A can be determined by comparing it with its translation in B. This disambiguation method has limited applicability, of course, because it requires a bilingual corpus. However, a corpus of translated text could be used to bootstrap catchword, providing it
with initial training instances (sentences containing a target homograph tagged with its sense), thereby eliminating the hand-labeling step.

Corpus-based methods are called ‘supervised’ when they learn from previously sense-annotated data, and, therefore, they usually require a large amount of human intervention to annotate the training data Ng (1997). Although several attempts have been made for example, Leacock (1998), Mihalcea (1999), Cuadros (2004), the knowledge acquisition bottleneck (too many languages, too many words, too many senses, too many examples per sense) is still an open problem that poses serious challenges to the supervised learning approach for WSD among Dravidian Languages including Tamil.

4.2.5 FINITE STATE TRANSDUCERS

A large part of the work on computational morphology has involved the use of finite-state devices, including the development of computational tools and infrastructure. Finite state methods have been used to handle both the strictly morphological and morphotactic states. We encounter the way in which strictly morphological information can be implemented with a finite state automaton, as in Figure 1.

By extending the notion of finite state automaton to that of finite state transducer, we can use much the same notions in order to generate the correct
surface morphemes and also create a device that can map surface sequences of letters (or phones) to abstract morphosyntactic features such as number, case and tense, aspect, mood etc. Computational morphology has also applied the notion of finite state transducer to deal with the problem of accounting for regularities of various sorts concerning alternative ways of realizing morphemes. For example, in Tamil, both the nominal suffix marking plural, case marker, clitic and the verbal suffix tense marking 3rd person singular is normally realized as -ai, -ukku, -āl for nouns -āy, -āl, -atu in the verbal inflection. We refer to this aspect of the problem as morphotactics, respectively. It would take us too far a field to go through an example in detail, but one could well imagine that the formation of the plural form of maram could be broken up into successive stages: maram + kaḷ → marañ+ kaḷ → marañkaḷ. Here, we see the suffixation of the plural ‘-kaḷ’ happening first, followed by the change of -am to -ā. In contrast, finite-state automata offer a way of dealing with the central phenomena of morphology without recourse to such a step-by-step inflection and derivation: hence the term ‘two-level morphology,’ which employs only two levels: one in which morphosyntactic features and lexical roots are specified, and one which matches the spelled (or pronounced) form of the word.

4.2.6 SPELL CHECKER

The purpose of spell checking is the detection and correction of typographic and orthographic errors in the text at the level of word occurrence considered out of its context. Nobody can write without any errors. Even people well acquainted with the rules of language can, just by accident, press a wrong key on the keyboard (maybe adjacent to the correct one) or miss out a letter. Additionally, when typing, one sometimes does not synchronize properly the movements of the hands and fingers. All such errors are called typos, or typographic errors. On the other hand, some people do not know the correct spelling of some words, especially in a foreign language. Such errors are called spelling errors.
First, a spell checker merely detects the strings that are not correct words in a given natural language. It is supposed that most of the orthographic or typographic errors lead to strings that are impossible as separate words in this language. Detecting the errors that convert by accident one word into another existing word, such as English then → ?than or Tamil palli (பல்லி) → ?palli (பல்லி), supposes a task which requires much more powerful tools.

After such impossible string has been detected and highlighted by the program, the user can correct this string in any preferable way - manually or with the help of the program. For example, if we try to insert into any English text the strings 3 *groop, *greit, or *misanderstand, the spell checker will detect the error and stop at this string, highlighting it for the user. Similarly if one types wrong spelling in Tamil, it can highlighting for the user *kaṭṭūrai (கடுளை), *aṉṆppur (அன்புப்பர்) *pallam (பல்லம்) can be highlighted it for the user.

The functions of a spell checker can be more flexible. The program can also propose a set of existing words, which are similar enough (in some sense) to the given corrupted word, and the user can then choose one of them as the correct version of the word, without re-typing it in the line. In the previous examples, Microsoft Word’s spell checker gives, as possible candidates for replacement of the string annan, the existing Tamil words shown in Figure-3.
In most cases, especially for long strings, a spell checker offers only one or two candidates (or none). For example, for the string *anuppugar, (அனுப்பகர்) it offers only the correct Tamil word anuppunar அனுப்பனர்).

The programs that perform operations of both kinds are called orthographic correctors, while in English they are usually called spell checkers. In everyday practice, spell checkers are considered very helpful and are used by millions of users throughout the world. In the modern text editors, if we take into consider the statistical study of Tamil morphology would certainly enhance the output of spell checker.

4.2.7 USE OF N-GRAM IN NLP

An n-gram is a stretch of text n character and words long. Information in n-grams tells us something about language, but doesn’t capture the structure. Finding and using every, e.g., two-word collocation in a text is quick and easy to do. An n-gram can help in a variety of NLP applications like character and word prediction and it can be used to aid in predicting the next character, word of an utterance, based on the previous n -1 words and also it is useful for context-sensitive spelling correction, approximation of language and so on. We make use of these corpora for n-gram with frequencies and their probability and percentage. Simple n-grams let us assume we want to predict the next word, based on the previous context of I dreamed I saw the knights in. What we want to find is the likelihood of w7 being the next word, given that we have seen w1, ...,w6, in other words, P(w1, ...,w7). In general, for w_n, we are looking for:

\[ P(w_1, ...,w_n) = P(w_1)P(w_2 | w_1)...P(w_n | w_1, ...,wn-1) \]

But these probabilities are impractical to calculate: they hardly ever occur in a corpus, if at all and it would be a lot of data to store, if we could calculate them.
4.3 APPLICATION TO SPEECH TECHNOLOGIES

Models of pronunciation variations are directly applicable to some branches of speech technologies, such as Text-to-Speech (TTS) systems and automatic speech recognition (ASR) systems. A TTS system, as the name suggest, convert written text to spoken speech. The system has many components, such as text analysis, duration prediction, and intonation prediction. Text analysis converts written text into phoneme string, duration prediction estimates how long each phoneme is, and intonation module predicts k=l contours. All relevant information are send to a speech synthesis component to produce speech output.

The text analysis component simulates the ability of an educated reader. Written text is analyzed and converted into a representation with morphological information, syllable structures, prosodic tags, phoneme strings and morpheme strings. Prediction of pronunciation variation is one area that presents interesting possibilities, but has hardly been explored in the TTS domain. This is primarily because the systems are still restricted to the reading mode.

Proper handling of pronunciation variations will be helpful in generating believable conversation speech, as well as simulating personal characteristics and regional accents. Speech recognition (ASR) is the reverse process of TTS. It takes speech as input and coverts it into text. TTS is an automatic reader, and ASR is an automatic transcriber. Pronunciation modeling is a crucial component of an ASR system. An ASR system evaluates acoustic signals against hypotheses of word pronunciation, based on transcriptions and pronunciation models. Multiple pronunciations of a word are annotated or generated so as to cover all the possible pronunciation variations. More complete annotation and good models of pronunciation variations can improve the performance of an ASR system. Pronunciation modeling of Tamil casual speech will be useful to integrate in ASR systems.
4.4 LANGUAGE TEACHING AND LEARNING

In language teaching and learning morphology plays an important role. Annotation of corpora is needed in sentence level as well as in word level. In an agglutinative language like Tamil the words are highly inflected with morphemes. For the word analysis we need to segment the morphemes for further study and tag its morphological categories. These grammatically annotated corpora are used for language teaching and learning. Tamil being a Dravidian language has a very rich morphological structure which is agglutinative, Tamil words are made up of lexical roots followed by one or more affixes. Lot of prior research has demonstrated that school and college level content words tend to be morphologically complex. Reading is the salient skill utilized across the curriculum and often the primary means of content dissemination. Reading, in turn, is principally linked to the extent of one’s vocabulary. Consequently, teaching morphologically complex vocabulary at the school and college preparatory level along with providing a working knowledge of morphemes can assist students toward school college readiness.

Corpora and the teaching syllabus

Large general corpora have proven to be an invaluable resource in the design of language teaching syllabi which emphasize communicative competence (cf. Hymes 1972, 1992) and which give prominence to those items that learners are most likely to encounter in real life communicative situations. In the context of computer corpus informed English language teaching syllabi, the first and probably most groundbreaking development was the design of the Collins COBUILD English Course (CCEC; Willis/Willis 1989), an offshoot of the pioneering COBUILD project in pedagogically oriented lexicography (cf. Sinclair 1987). The contents of this new, corpus-driven “lexical syllabus” are “the commonest words and phrases in English and their meanings” (Willis 1990). With its focus on lexis and lexical patterns, the CCEC responds to some
of the most central findings of corpus research, namely that language is highly patterned in that it consists to an immense degree of repeated word-combinations, and that lexis and grammar are inseparably linked (cf. Hoey 2000; Hunston 2002; Hunston and Francis 2000; Partington 1998; Römer 2005a, 2005b; Sinclair 1991; Stubbs 1996; Tognini Bonelli 2001). Also worth mentioning is a much earlier attempt to improve further the teaching of English vocabulary that was made long before the advent of computers and electronic corpora. In 1934, Michael West organized a conference “to discuss the part played by corpus-based word lists in the teaching of English as a foreign language” (Kennedy 1992, 327). About 20 years later, West’s (1953) General Service List of English Words (GSL) was published and has since then exerted great influence on curriculum design (cf. Kennedy 1992, 328; Willis 1990, 47). As the title indicates, West’s GSL suggests a syllabus that is based on words rather than on grammatical structures. It is also based on frequently occurring rather than on rare words. Of course, frequency of occurrence is not the only criterion that should influence decisions about the inclusion of items in the teaching syllabus (there are other relevant criteria, such as “range, availability, coverage and learnability (Mackey 1965, 188)” (Kennedy 1992, 340); cf. also Nation 1990, 21), but it is certainly an immensely important one (see also Aston 2000, 8; Leech 1997, 16). It can be safely assumed that learners will find it easier to develop both their receptive and productive skills when they are confronted with the most common lexical items of a language and the patterns and meanings with which they typically occur than when the language teaching input they get gives high priority to infrequent words and structures which the learners will only rarely encounter in real-life situations.

Another thread in applied corpus research that aims to inform the teaching syllabus and also stresses the importance of frequency of occurrence, examines language items in actual language use and compares the distributions and patterns found in general reference corpora (of speech and
writing) with the presentations of the same items in teaching materials (course books, grammars, usage handbooks). The starting point for these kinds of studies is usually language features that are known to cause perpetual problems to learners, for example, for German, discourse particles (Jones 1997), modal verbs (Jones 2000), the passive voice (Jones 2000) prepositions (Jones 1997), or, for English, future time expressions (Mindt 1987, 1997), if-clauses (Römer 2004b), irregular verbs (Grabowski/Mindt 1995), linking adverbials (Conrad 2004), modal verbs (Mindt 1995; Römer 2004a), the present perfect (Lorenz 2002; Schlüter 2002), progressive verb forms (Römer 2005a, 2006) and reflexives (Barlow 1996). For all these phenomena, researchers have found considerable mismatches between naturally-occurring German or English and the type of German or English that is put forward as a model in the examined teaching materials. They have, as a consequence, called for corpus-inspired adjustments in the language teaching syllabus (particularly as far as selection and progression are concerned) and for revised pedagogical descriptions which present a more adequate picture of the language as it is actually used. A case in point here is the misrepresentation of the functions and contextual patterns of English progressive forms in EFL teaching materials used in German schools. Progressives that refer to repeated actions or events, for example, are considerably more frequent in ‘real’ English than in textbook English where the common function “repeatedness” is rather neglected and the focus is on single continuous events (cf. Römer 2005a, 261-263).

As Ellis point out us that language learners naturally rely on frequency for many different language tasks, ranging from irregular verb patterns to collocation patterns. Given its importance in acquisition, we would argue that frequency should also play a key role in the development of materials and in the choices that teachers make in language classrooms. With the recent availability of comprehensive frequency-based grammatical descriptions, such integration of pedagogy and research has become feasible. Actually,
textbook authors and students have been interested in frequency information all along. For example, many books include lists of common phrasal verbs or provide other information about patterns that are common or typical. However, more often than not, this information is based on the author’s intuitions rather than empirical research. As a result, these lists often include forms that are not in fact common, while overlooking other forms that do occur frequently. (Surprisingly, extremely frequent forms are also the ones that we are most likely to overlook.) By using information based on actual frequency and context of use (e.g., register differences), materials developers and teachers should be able to increase the meaningful input that is provided to learners. Obviously, other factors are equally important - for example, some grammatical topics are required as building blocks for later topics; some grammatical topics are more difficult and therefore require more practice than others. In many cases, though, language use has been a primary guiding principle: Authors have attempted to present the typical and most important patterns first, moving on to more specialized topics in later chapters and more advanced books. Lacking empirical studies, authors have been forced to rely on their intuitions for these judgments about language use, and widely accepted norms have arisen to support those intuitions. With the rise of corpus-based analysis, we are beginning to see empirical descriptions of language use, identifying the patterns that are actually frequent (or not) and documenting the differential reliance on specific forms and words in different registers. In some cases, our intuitions as authors have turned out to be correct; in many other cases, we have been wrong. For those latter cases, revising pedagogy to reflect actual use, as shown by frequency studies, can result in radical changes that facilitate the learning process for students.

In this study, I wish to examine the relationship between corpus linguistics (CL) and language teaching (LT) and provide an overview of the most important pedagogical applications of corpora. As the below figure aims to illustrate, this relationship is a dynamic one in which the two fields greatly
influence each other. While LT profits from the resources, methods, and insights provided by CL, it also provides important impulses that are taken up in corpus linguistic research. The requirements of LT hence have a pact on research in CL and on the development of suitable resources and tools. This kind of study would investigate what influence CL has had on LT so far, and in what ways corpora have been used to improve pedagogical practice. It will also discuss further possible effects of CL on LT and of LT on CL, and highlight some future tasks for researchers and practitioners in the field.

4.5 SPEECH RECOGNITION AND SYNTHESIS

Both applications can benefit from morphological information. As Demberg et al. (2007) showed that morphological segmentations provided by rule-based systems improve the grapheme to phoneme conversion process. For speech recognition in morphologically complex languages such as Finnish and Tamil and other Dravidian Languages, language models based on morphs instead of words are more robust with respect to the well-known problem of out of vocabulary words (Creutz et al., 2007).

4.6 TERMINOLOGY

Most of the work in terminological data processing focuses on the detection of lexical variants in complex terms, in order to fine-tune controlled term indexing and term extraction. Amongst the lexical phenomena that make a
term vary in context, morphology (both inflectional and constructional) plays an important role (Jacquemin et al., 1997).

4.7 TEXT CLASSIFICATION

Linguistic units which are morphologically motivated can be used as features for text categorization, leading to a reduction of the dimension of the feature vectors and improved performance for a language with many compounds such as German (Witschel and Biemann, 2006). Thus, statistical study of morphology is useful for a large variety of NLP applications.

4.8 CONCLUSION

The statistical study of morphology on corpus of Tamil language has been used to improve dramatically in the last decades. These improvements had a positive effect on the computational linguistics areas. We have seen that how these NLP applications such as morphological analyzer, morphological generator, machine translation, parts-of-speech, word sense disambiguation and spell checker etc, are made use of statistical study of Tamil morphology which is very importance to enhance these systems.