Inter-Language Analysis of POS-Tagging

VIKTOR EKLUND
and FREDRIK LINDH

KTH Computer Science
and Communication

Bachelor of Science Thesis
Stockholm, Sweden 2010

Inter-Language Analysis of POS-Tagging

VIKTOR EKLUND
and FREDRIK LINDH

Bachelor's Thesis in Computer Science (15 ECTS credits)
at the School of Computer Science and Engineering
Royal Institute of Technology year 2010
Supervisor at CSC was Johan Boye
Examiner was Mads Dam

URL: www.csc.kth.se/utbildning/kandidatexjobb/datateknik/2010/
eklund_viktor_OCH_lindh_fredrik_K10073.pdf

Royal Institute of Technology
School of Computer Science and Communication

KTH CSC
100 44 Stockholm

URL: www.kth.se/csc

Referat

Den här rapporten utforskar skillnaderna mellan olika Part-Of-Speech Taggers (POST) applicerat på olika språk. Åtminstonde två olika implementationer av POST som bygger på markovmodeller jämförs och körs sedan på korpus av liknande typer, varpå utdata jämförs med korrekt annoterade facit. Felen som identifieras undersöks och klassificeras efter typ. I den påföljande analysen undersöks om felen är konsekventa inom ett givet språk samt om mönster bland felen kan ses mellan språken. Till sist hålles en diskussion relevant till utdatan och analysen där förslag till ändringar och tillägg till algoritmerna / programmen föreslås.

Abstract

This report explores the differences between different Parts-Of-Speech Taggers (POST) in different languages. At least two different implementations of POST using Markov models will be compared and then executed with corpuses of the same kind. The resulting data will then be compared with correctly annotated text. The identified errors will then be examined and classified by type. Following that, an analysis is conducted to determine whether the errors follow a pattern within a language, and if the same errors appear consistently between different languages. Finally a concluding chapter will bring together the data acquired from the tests with suggestions to possible amendments and additions to the algorithms the tagging software uses.

Contents

1 Introduction 5

1.1 Background 5
1.2 Problem statement 5
1.3 Brief explanation of parts of speech 5
1.4 Brief explanation of a corpus 5

2 Relevant tagging tools 6

2.1 MeCab...... 6

2.2 Stanford POST 7

2.3 MXPOST 8

2.4 Brief explanations of concepts used by the taggers 8
2.4.1 Maximum Entropy model 8
2.4.2 Viterbi Algorithm 8
2.4.3 Hidden Markov Models 8

3 Relevant corpuses 9

3.1 Open American National Corpus 9

3.2 Stockholm Umeå corpus 2.0 9

3.3 Similarities and differences 9

4 Method 10

4.1 Corpus tagging 10

4.1.1 Hypothesis 11
4.1.2 Execution of the english part 12
4.1.3 Results 13
4.1.4 Execution of the japanese part 17
4.1.5 Results 18
4.1.6 Japanese tounge twisters 27

5 Conclusions 28
6 Bibliography 29
Appendix A 30
Appendix B 31
Appendix C 32
Appendix D 37

1 Introduction

1.1 Background

Part-Of-Speech Tagging (Henceforth referred to as POST or POS tagging) has a long history in the area of speech/text-analysis, implementation of expert systems and ELIZA[1] type consoles. A large number of different algorithms are currently in use, most of them designed for and applied to English. It is also the most widely used basis when delving into the area of semantic parsing which is one of the major precursors to making AI interfaces[16], as well as being the core in a set of advanced web based query systems[15]. It is therefore imperative that the POSTs used are working at optimum efficiency. There has been a lot of study in the sub-discipline of comparing and optimizing different taggers operating at the same language, however this is not true in the case of cross-language analysis. We have in light of this decided to engage in a study consisting of three different languages (Swedish, English and Japanese) parsed with markov model-powered POST algorithms.

1.2 Problem Statement

Our aim with this study is to identify patterns of failure for a POST and once, assuming that we do, identify such a pattern, our goal will be to suggest and/or implement a solution. To help us identify such patterns we will be using at least two already established POS taggers for different languages, the reason being that patterns of failure could possibly be different given differences in grammatical structure of two different languages, but with the same algorithms implemented. Our method of choice for said taggers will be the Maximum Entropy model, which is based on hidden Markov models.

1.3 Brief explanation of parts-of-speech tagging

Parts-of-speech tagging is the process of assigning to each word in a text, their corresponding part of speech. The Pen Treebank tag set in appendix A serves as a list of the possible part of speech tags in the English language. For example, if we were to tag the sentence “Let’s eat!” it would (hopefully) come out as “Let_VB 's_PRP eat_VB !_.” From these tags we’re now able to see that, in this particular sentence, ‘Let’ is considered to be a verb, “ ’s ” is considered a preposition, e.t.c. In computer science, the challenge is then to design and implement an algorithm that, as correctly as possible, can assign tags to a corpus in reasonable amounts of time.

1.4 Brief explanation of a corpus

A corpus is simply a large collection of texts, often categorized into their respective types, such as fiction, non-fiction, travel guides, scientific, e.t.c. Different uses for a corpus include, as we will be doing, tagging them to evaluate the efficiency of the implemented algorithm. A tagged corpus can be of greater use for linguists than an untagged one.

2 Relevant tagging tools

2.1 MeCab

MeCab [4] is an open source POST for Japanese. It is built upon the earlier tagger “Chasen” and originally shared a lot of its structure and functions (as ”ChasenTNG”). But the original program was abandoned and its creator Taku Kudou eventually recreated it completely from scratch. Being approximately 4 times faster than Chasen, with the same level of results, it is currently considered one of, if not the best Japanese POST available. Some of MeCab's features include:

- Not corpus reliant
- Uses a concept similar to markov random fields as its core called CRFs
- Easy to refactor with an easy to understand open source library
- Implementation friendly (SDKs for JAVA, PERL, PYTHON, RUBY & C#)

While it uses bi-gram markov models for analysis (which is the theme for the taggers in this thesis), its learning algorithm is based on the not so common CRF method. CRF stands for Conditional Random Fields and is similar to a hidden markov model. It is represented by an undirected graphical model in which each vertex represents a word whose tag is to be inferred, and each edge represents a dependency (statistical or environmental) between two vertices. These vertices and edges, like sentences, form chains rather than interconnected graphs, and are then used to sequentially tag the input.

As most other POSTs it uses the Viterbi search algorithm to construct its output (most probable path) and is constructed to be able to parse an infinite number of possible paths at the same time. It handles unknown words (words not found in the training data) by assigning the most probable word class to them as inferred by its position and possibly surrounding words. The most common word classes assigned to unknowns are the classes noun and verb, although the weights controlling this behavior are configurable. MeCab also has a few other features such as the N-Best feature and numerous other runtime flags. While performing the Viterbi analysis, MeCab keeps track of secondary solutions in ranked order down to the N-Best solution. By modifying a runtime flag the user is able to produce up to an unlimited number of secondary solutions if necessary. Other flags of interest are those controlling input/output piping for easier handling of larger texts, flags controlling encoding settings and flags for different types of training rules.

A big difference between western and eastern POST is the problem of sentence splitting (commonly referred to as ”wakachi” 「分かち」 in the Japanese POST community), which requires a lot of effort, especially in Japanese and Chinese. Sentence splitting is essentially the problem of dividing a sentence into tokens. While western language ”gets this for free” due to our habit of using delimiter characters between almost every word, some Eastern languages do not, and consequently have to solve that problem before the normal POS tagging can begin.

2.2 Stanford POST

For the purpose of this investigation we will be using the Stanford "Log-linear" Parts-of-speech tagger[7]. The software uses a Maximum Entropy model for tagging text with the Penn Treebank POS tag set. The tagger has been implemented to support any language, provided the tagger is supplied with a tagging model file for that language, and that the language is white space separated.

If such a model file does not exist, or if one simply wants to create a new tagging model for a language for which a tagging model already exists. One could easily create one by training the tagger on correctly tagged example text. We will not be interested in this feature as a highly trained English model is already provided, and for the Swedish part of this study we've decided to use another tagger.

The Stanford tagger is licensed under the GNU full GPL[3], which allows it to be used for research purposes and free software projects. It provides complete javadocs and source code for easy implementation and customization to any kind of non commercial and non proprietary software project wanting to use it.

To run the Stanford tagger in a java application one simply instantiates the maximum entropy tagger with

```
MaxentTagger tagger = new MaxentTagger();
```

With an option overloaded constructor available for a parameter to a different tagging model than the default English tagging model. After that, the file to be tagged can be loaded and whitespace tokenized with

```
List<Sentence<? extends HasWord>> sentences =
MaxentTagger.tokenizeText(new BufferedReader(new FileReader(“fileName”)));
```

Lastly, one can tag the loaded text and write the result to a file with

```
BufferedWriter fileWriter = new BufferedWriter(new FileWriter(“outputFileName”));
for(Sentence sentence : sentences){
Sentence<TaggedWord> tSentence = MaxentTagger.tagSentence(sentence);
    fileWriter.write(tSentence.toString(false));
}
```

Naturally, the tagger provides many other interfaces for tagging text as well as, like mentioned earlier, methods for training the tagger and testing accuracy. It can also, if one wishes, be run directly from the command line, without losing access to any functionality.

2.3 MXPOST

MXPOST (MaXimum Entropy Part-Of-Speech Tagger) is a tagger originally written by Adwait Ratnaparkhi in 1996 to accommodate English, but has since its creation been adapted to a number of languages, among them Swedish and Dutch. It is a also a component in a number of software applications, like the MARY Text-To-Speech system.

As the name suggests, MXPOST, like the Stanford tagger, is based on maximum entropy learning algorithms and uses beam search to find the most probable sequence of tags. Like other probabilistic taggers, MXPOST analyzes a text by inspecting relations between tokens and also by checking the first and last letters in a given word.

2.4 Brief explanations of concepts used by the taggers

2.4.1 Maximum Entropy Model

Is a technique for estimating input probabilities of a process that is consistent, with known constraints expressed in terms of averages, or expected values, of one or more quantities, but is otherwise as unbiased as possible.
This means that all states in the model initially have the same probability of occurrence, which are then modified by various constraints imposed on the model. These constraints then let us infer data from the model by the use of various algorithms.

2.4.2 Viterbi Algorithm

The Viterbi algorithm is used to find the most likely sequence of hidden states. The result of the algorithm is called the ”Viterbi path” and consists of a series of observed events, which makes it a type of hidden markov model. The algorithm is usually supplied with a set of states and a set of rules describing the probability that two states are arranged in a certain order. The algorithm propagates forward from the first instance in the set and selects the highest cumulative probability, relative to the current path, as it's child until the algorithm reaches the target instance in the set. When this occurs the algorithm back propagates, marking the correct path in inverse order.

2.4.3 Hidden Markov Models

A hidden markov model is a statistical model where the states in the model are invisible to the observer except for the output. Each state has a set of probabilities mapped to a set of possible output tokens. Because of this, the sequence of states can to some extent be inferred from observations of the model.

3 Relevant corpuses

3.1 Open American National Corpus

The Open American National Corpus[2] (henceforth OANC) is a freely distributed sub corpora of the more rigorously tagged ANC "Second release", containing more than 14 million words. It is categorized in the different kinds of texts that it's composed of, categories such as travel guides, journals, fictional texts and letters. For our investigation we will concern ourselves with the fictional category.

The OANC is available with POS tags according to the Penn Treebank tag set, which has first been automatically generated and corrected afterwards[2]. As such, the tags provided in the OANC corpus are not guaranteed to be 100% correct (and indeed, no corpora of a significant size are). It will however serve its purpose just as well.

3.2 Stockholm Umeå corpus 2.0

From the SUC homepage [8] :

The Stockholm Umeå Corpus is a Swedish corpus of 1 million words, in which each word has been tagged, i.e. annotated with its part-of-speech, inflectional form and lemma. All the texts in the corpus were written in the 1990's, and are balanced according to genre, following the principles used in the Brown and LOB corpora. SUC was developed in a joint project between the universities of Stockholm and Umeå, and it is freely distributed for research purposes.

The first task was acquiring the corpus, which entailed getting the research license. This delay is one of the causes contributing to the failure of the Swedish part of the project. As soon as we had acquired the license, we began by examining the corpus. The SUC uses the “SUC Morphosyntactic tag set” which is based on the SWETWOL[10] tag set. It is similar to the Parole tagset[13] and a one to one mapping between them exists[11]. The tags are produced by concatenating a POS tag with a series of word feature tags. The resulting tags are quite verbose compared to other tag sets.

3.3 Similarities and differences

Differences between the two corpuses that we will use obviously need to be presented and taken into account. The first thing to note is that the two corpuses are annotated with two different tag sets. The Stanford POST uses the Penn Treebank tag set whereas the Stockholm Umeå corpus uses, as stated above, their own “Morphosyntactic tag set”. This did not worry us much however, as mostly different tag sets simply look different but represent the same things.

Fortunately, both of the corpora are categorized into different kinds of text. We’ll pick a fictional literature category, which is available from both corpora, in order to avoid results that might be heavily influenced by the way different text categories are tagged.

Differences in the accuracies of the different corpora is not of much importance, as this investigation will be interested in cases where there is a tagging conflict between what the tagger software suggests, and the “correct” tags provided by the corpus, and corpus annotation accuracy does not need to be 100% for this to yield interesting results.

4 Method

4.1 Corpus tagging

To process the tagging output, several tools had to be programmed. First of all a text pre-processor had to be designed to remove certain features of the input files that produced errors when tagged in the form of out of bound characters. The next step was to write a post-processor to filter excessive whitespace and other scrap information from both the tagger output and the benchmark. Spacing was also added around symbol characters to make tokenizing easier. As the benchmark corpus' tagging convention was slightly different from the taggers output, filters had to be added to counteract this. After filtering, words were tokenized and line breaks were added after each token for further processing. Finally, a manual parser was written to allow us to ensure that the files were identical as far as tokens and formatting went. Some artifacts that were not caught by the filters could also be weeded out in this step. Once the texts had been through processing the problem turned to producing a parser that would go through the two texts and generate whatever statistical information we might want. The programs were all written in C# with Windows Forms, and the source code for all the projects can be supplied on request.

The MXPOST program on the other hand presented a multitude of problems. Because of the lack of documentation and unintuitive interface, setting up the program took several hours. Initial testing with the included project went relatively fine, reporting approximately 94% accuracy on circa 10.000 words. The next step; training of the Swedish model is where things started to go awry. We started by running the trainer on a training set of 10.000 words. The training session took about one day to complete. A model was indeed produced, but the accuracy was abysmal, only about 5% of the words were annotated correctly when running tests on sentences about 20 words long. Most of the tags used were also not valid tags at all. We assumed that the non-valid tags were some type of default tag and attributed the accuracy to the small training set. We then decided to try to tag a larger model, with hopes of better results. We initiated one training session on a 1Ghz, 512MB Unix laptop with a 250.000 word training set, and another on a quad core 2.6Ghz 4096MB Windows machine using a 1.000.000 word training set. After two days, the 250k model completed with very strange results. The accuracy was still only about 5%, with the same non-existent tags saturating the output. An additional test was run on a smaller training set while waiting for the 1000k session but it resulted in similar errors. We theorize that the MXPOST tagger is not compatible with any other type of tag syntax than the Penn Treebank set, and since the SUC uses the "SUC Morphosyntactic tag set" it produced corrupt models. A solution to this problem would be to write a custom translating tool between the two sets and run the tagger on the output. This, however, with the time available is quite an insurmountable task and might constitute a project in itself. When the 1000k finally terminated, the model produced corrupt output as well, as expected. Because of these problems, we had to scrap the Swedish part of the project. As such, research into the relative errors in tags between linguistically similar languages (e.g. English and Swedish) might still be warranted

The Stanford POST package however, as mentioned earlier, came with an already highly trained English model. This saved us the time of having to locate training data and training the tagger, a process which could easily take a couple of days. Using the tagger also proved very straightforward and we quickly had one version of the text tagged by the Stanford POST and one version with the provided OANC tags.

4.1.1 Hypothesis

Our initial hypothesis was that idioms and figures of speech would be the most likely to have erroneous tags. We also considered highly ambiguous sentences like the classic example “Time flies like an arrow”, which has a total of (at least) five different interpretations [12]

One could theorize that a statistical tagger like the ones we’ll be testing could care less about ambiguous sentences, seeing how ambiguity is mostly a problem for human readers, and a correctly annotated training set for the tagger would make it so that statistics gets it right anyway.

We held on to this belief however as there is no such thing as correctly annotated training data, even the best training data available has a certain percentage of erroneous classifications and ambiguous sentences are the most likely to have been miss-tagged by the people that create the training data, as of course, it has to be tagged and created by hand.

Another reason that we held on to this belief is that words that are part of ambiguous sentences are more likely to have several possible tags depending on their context, which would make statistics for these words less accurate.

Another assumption we had at the early stages of our investigation was that we truly expected us to see error patterns between the different languages, and that said patterns would lead to heuristics for how to better address the most common error patterns identified.

These were our hypotheses and thoughts as we started investigating the results and numbers presented below this section, and generated as described in the section above.

4.1.2 Execution of English part

The Stanford POST was run on the largest single text file of the OANC corpus, half a megabyte of text consisting of mostly an excerpt from a novel, but also featuring some other strange text pieces like cooking recipes. This was done because of our initial intention to eventually analyze a set of sentences and this amount of data provided a satisfactory amount of erroneous sentences. The resulting tags were compared in various ways to the tags provided by the OANC to generate statistics. It was a simple enough task to determine that the tags matched 94% of the time.

The information that appealed the most to us and which was also the most relevant to our goal however were the numbers shown when we decided to calculate frequencies of error combinations. Below is a table showing the frequency of error combinations in percent of the total number of errors for the combinations that were responsible for at least 4% of all errors. We also decided that, seeing how the distributions were fairly equal, if an error was of the kind X in one text and Y in the other, it would count the same as the opposite occurrence. We also did not distinguish between known and unknown words.

Tag combinations	**Percent of total errors**
JJ/NN	13.56 %
NNP/NN	5.90 %
JJ/NNP	5.36 %
VBD/VBN	4.92 %
VBP/VB	4.86 %
VBN/ JJ	4.43 %
NN/VB	4.40 %
VBZ/NNS	4.29 %

In light of this result and seeing how the three most common misclassifications are essentially of the same kind, our decision was to extract and analyze some of the sentences that had misclassifications of the types JJ/NN, VBZ/NNS and VBN/JJ

4.1.3 Results

Sentence	An evil young girl pacified a colicky toddler with wine spritzers in her baby bottle.
OANC Tags	An_DT evil_JJ young_JJ girl_NN pacified_VBD a_DT **colicky_NN** toddler_NN with_IN wine_NN spritzers_NNS in_IN her_PRP$ baby_NN bottle_NN ._.
Stanford Tags	An_DT evil_JJ young_JJ girl_NN pacified_VBD a_DT **colicky_JJ** toddler_NN with_IN wine_NN spritzers_NNS in_IN her_PRP$ baby_NN bottle_NN ._.
Comment	The correct interpretation is the Adjective (JJ) one, as colicky in this sentence is a disease afflicting the baby. The erroneous OANC tag might be because of bad initial tagging or because the tagger recognized ”colicky toddler” as one word. Another possible source might have been that the word was not in the training set and was tagged as NN as default.

Sentence	The sad-ness itself will already by an explanation of a somatic sensation or mental phenomenon.
OANC Tags	The_DT sad-ness_JJ itself_PRP will_MD already_RB be_VB an_DT explanation_NN of_IN a_DT somatic_JJ sensation_NN or_CC mental_JJ phenomenon_NN ._.
Stanford Tags	The_DT sad-ness_NN itself_PRP will_MD already_RB be_VB an_DT explanation_NN of_IN a_DT somatic_JJ sensation_NN or_CC mental_JJ phenomenon_NN ._.
Comment	The correct interpretation is the Noun(NN) one, as it is in fact equivalent to ”sadness”. The error might have occurred due to the hyphen, which makes the tagger interpret the word a compound adjective, e.g. ”cool-ness” or ”sweet-ness”. This is a good example of when bad formatting might influence the output.

Sentence	The tree, as the rule of the physical universe would have it, reflects light.
OANC Tags	The_DT tree_NN ,_, as_IN the_DT rule_NN of_IN the_DT physical_JJ universe_NN would_MD have_VB it_PRP ,_, reflects_VBZ **light_JJ** ._.
Stanford Tags	The_DT tree_NN ,_, as_IN the_DT rule_NN of_IN the_DT physical_JJ universe_NN would_MD have_VB it_PRP ,_, reflects_VBZ **light_NN** ._.
Comment	The correct interpretation is the Noun(NN) one, as it is the primary object of the sentence. The error is probably due to bad parsing of the sentence. The somewhat sparse secondary clause might have affected the probability series as it starts with a verb, and in this situation the previous sentence structure in addition to the secondary clause might have promoted the erroneous choice.

Sentence	He gave Blanche the water-color sketch.
OANC Tags	He_PRP gave_VBD Blanche_NNP the_DT **water-colour_JJ** sketch_NN ._.
Tagged as	He_PRP gave_VBD Blanche_NNP the_DT **water-colour_NN** sketch_NN ._.
Comment	The correct interpretation is the Noun (NN) one, as it is a compound word together with ”sketch” and not an attribute. The hyphen might have been a significant factor in this instance as well. But the more probable cause is the DT-X-NN pattern which occurs at the end of the sentence. In this case, a DT-JJ-NN pattern is much more probable than a DT-NN-NN one. This showcases one of the primary weaknesses of statistical tagging.

Sentence	You have eyes like a Siberian Husky; underneath the red they are ice blue.
OANC Tags	You_PRP have_VBP eyes_NNS like_IN a_DT Siberian_NNP Husky_NNP ;_: underneath_IN the_DT red_JJ they_PRP are_VBP ice_NN **blue_JJ** ._.
Stanford Tags	You_PRP have_VBP eyes_NNS like_IN a_DT Siberian_NNP Husky_NNP ;_: underneath_IN the_DT red_JJ they_PRP are_VBP ice_NN **blue_NN** ._.
Comment	The correct interpretation is the Noun(NN) one, as it is a compound word together with ”ice”. The error probably occurred because there are very few instances of the word ”blue” being treated as a noun as it is in this compound, and there are no real indicators of the words compound nature.

Sentence	I liked that little recitative by Ilia in Idomeneo.
OANC Tags	I_PRP liked_VBD **that_IN little_RB recitative_JJ** by_IN Ilia_NNP in_IN Idomeneo_NNP ._.
Stanford Tags	I_PRP liked_VBD **that_DT little_JJ recitative_NN** by_IN Ilia_NNP in_IN Idomeneo_NNP ._.
Comment	The correct interpretation is DT-JJ-NN, as it is an event hosted by ”Ilia”. It is hard to discern the reason for this error, but one could surmise that ”that little” got grouped because of the initial part of the sentence. In the context, the probability for the conjunction tag might have been higher, and the other errors might have come as a result of the initial error which produced a faulty set of sentence features.

Sentence	Here is the face of a patient, drugged out of pain and sorrow, drifting on Demerol.
OANC Tags	Hers_JJ is_VBZ the_DT face_NN of_IN a_DT patient_NN ,_, **drugged_JJ** out_IN of_IN pain_NN and_CC sorrow_NN ,_, drifting_VBG on_IN Demerol_NNP ._.
Tagged as	Hers_JJ is_VBZ the_DT face_NN of_IN a_DT patient_NN ,_, **drugged_VBN** out_IN of_IN pain_NN and_CC sorrow_NN ,_, drifting_VBG on_IN Demerol_NNP ._.
Comment	The correct interpretation here is ambiguous, it is unclear if it is being emphasized that ”She” was drugged, in which case the correct tag would be Verb(VBN) or if it is simply a description of her state in the main clause, in which case the correct tag would be Adjective(JJ).

Sentence	That baby was nearly electrocuted in the Casino.
OANC Tags	That_DT baby_NN was_VBD nearly_RB **electrocuted_JJ** in_IN the_DT Casino_NNP ._.
Tagged as	That_DT baby_NN was_VBD nearly_RB **electrocuted_VBN** in_IN the_DT Casino_NNP ._.
Comment	The correct interpretation is the Verb(VBN) one, as it describes a possible event with a passive action. The error is probably due to the VBD-RB-X part of the sentence. A verb rarely follows a verb->adverbial.

Sentence	Somebody has come into camp with a stringed instrument and is singing about Divine Love in Provencal.
OANC Tags	Somebody_NN has_VBZ come_VBN into_IN camp_NN with_IN a_DT **stringed_JJ** instrument_NN and_CC is_VBZ singing_VBG about_IN Divine_NNP Love_NNP in_IN Provencal_NNP ._.
Tagged as	Somebody_NN has_VBZ come_VBN into_IN camp_NN with_IN a_DT **stringed_VBN** instrument_NN and_CC is_VBZ singing_VBG about_IN Divine_NNP Love_NNP in_IN Provencal_NNP ._.
Comment	The correct interpretation is the Adjective(JJ) one, as there is no action involved, it is simply a description (attribute) of the instrument. The previous tag sequence is probably the source for error here as well, although we imagine the probabilities were closer between the two tags in this case compared to the earlier ones.

Sentence	Our own barbarians vs theirs now!
OANC Tags	Our_PP$ own_JJ barbarians_NNS **vs_NNS** theirs_PRP now_RB !_.
Tagged as	Our_PP$ own_JJ barbarians_NNS **vs_VBZ** theirs_PRP now_RB !_.
Comment	Both tags are wrong as vs is equivalent to ”versus” which is a Preposition(IN). The most probable reason for error is probably that the word was incorrectly tagged in the training data, and being a rare word might have defaulted to one generally unlikely tag because it was one of the very few uncertain tags available.

Sentence	Siegfried sniffles as his eyes refocus on the Autobahn.
OANC Tags	Siegfried_NNP **sniffles_NNS** as_IN his_PP$ eyes_NNS refocus_VB on_IN the_DT Autobahn_NNP ._.
Tagged as	Siegfried_NNP **sniffles_VBZ** as_IN his_PP$ eyes_NNS refocus_VB on_IN the_DT Autobahn_NNP ._.
Comment	The correct interpretation is the Verb(VBZ) one, as Siegfried is taking an action. A very strange tagging error. In addition to the NNP-NNS-IN sequence being very unlikely, the VBZ tag should be more common for the word in general. One explanation might be that the training data was skewed in favor of the generally less common NNS tag, and therefore tagged it as such because of the skewed set.

Sentence	Caprice worries about her husband.
OANC Tags	Caprice_NNP **worries_NNS** about_IN her_PP$ husband_NN ._.
Stanford Tags	Caprice_NNP **worries_VBZ** about_IN her_PP$ husband_NN ._.
Comment	The correct interpretation is the Verb(VBZ) one, as Caprice is taking an action. This is another example of the NNP-NNS-IN sequence. The fact that there are more of these examples is very interesting. It suggests that there might have been some kind of logical error in the algorithm or very biased training data for some parts of the corpus.

4.1.4 Execution of Japanese part

After completing the stanford tagging experiment we moved on to phase two; comparing our tagging output data to the output of the erring sentences translated to Japanese. The translations were made by Lindh and proofread by a Japanese contact to ensure complete accuracy. The translations themselves are semantic translations, and not very litteral as features in the languages might impact the tagging results. The words with conflicting tags however, was translated in an similar or identical context to make the comparison viable. The tagger used was MeCab with the provided training model (ipadic). Installation and training went very smoothly compared to mxpost and tagging could commence almost immediately. The results for the translated control sentences were very interesting indeed and was not really what we anticipated. They are as follows:

4.1.5 Results

English Sentence	An evil young girl pacified a **colicky** toddler with wine spritzers in her baby bottle.
Japanese Sentence	邪悪な女が**夜鳴きする**赤ん坊をほ乳瓶に入ったワインであやしている。
MeCab Output	邪悪 名詞, 形容動詞語幹, *, *, *, *, 邪悪, ジャアク, ジャアク な 助動詞, *, *, *, 特殊・ダ, 体言接続, だ, ナ, ナ 女 名詞, 一般, *, *, *, *, 女, オンナ, オンナ が 助詞, 格助詞, 一般, *, *, *, が, ガ, ガ **夜鳴き** 名詞, 一般, *, *, *, *, 夜鳴き, ヨナキ, ヨナキ **する** 動詞, 自立, *, *, サ変・スル, 基本形, する, スル, スル 赤ん坊 名詞, 一般, *, *, *, *, 赤ん坊, アカンボウ, アカンボー を 助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ ほ乳 名詞, サ変接続, *, *, *, *, ほ乳, ホニュウ, ホニュー 瓶 名詞, 一般, *, *, *, *, 瓶, ビン, ビン に 助詞, 格助詞, 一般, *, *, *, に, ニ, ニ 入っ 動詞, 自立, *, *, 五段・ラ行, 連用タ接続, 入る, ハイッ, ハイッ た 助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ ワイン 名詞, 一般, *, *, *, *, ワイン, ワイン, ワイン で 助詞, 格助詞, 一般, *, *, *, で, デ, デ あやし 動詞, 自立, *, *, 五段・サ行, 連用形, あやす, アヤシ, アヤシ て 助詞, 接続助詞, *, *, *, *, て, テ, テ いる 動詞, 非自立, *, *, 一段, 基本形, いる, イル, イル 。 記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (noun-verb). The adjective in the English sentence turns into a noun verb combination in Japanese. The sino-japanese compound effectively prevents mistagging as the noun and verb are completely separated. The describing meaning of "colicky" is in other words translated to "who cries at night".

English Sentence	The **sad-ness** itself will already by an explanation of a somatic sensation or mental phenomenon.
Japanese Sentence	この**かなーしみ**そのものが既に病人の肉体的感覚と精神的現象を暗示している。
MeCab Output	この　連体詞, *, *, *, *, *, この, コノ, コノ **か**　助詞, 副助詞／並立助詞／終助詞, *, *, *, *, か, カ, カ **なー**　助詞, 終助詞, *, *, *, *, なー, ナー, ナー **しみ**　名詞, 一般, *, *, *, *, しみ, シミ, シミ そのもの　名詞, 一般, *, *, *, *, そのもの, ソノモノ, ソノモノ が　助詞, 格助詞, 一般, *, *, *, が, ガ, ガ 既に　副詞, 一般, *, *, *, *, 既に, スデニ, スデニ 病人　名詞, 一般, *, *, *, *, 病人, ビョウニン, ビョーニン の　助詞, 連体化, *, *, *, *, の, ノ, ノ 肉体　名詞, 一般, *, *, *, *, 肉体, ニクタイ, ニクタイ 的　名詞, 接尾, 形容動詞語幹, *, *, *, 的, テキ, テキ 感覚　名詞, 一般, *, *, *, *, 感覚, カンカク, カンカク と　助詞, 並立助詞, *, *, *, *, と, ト, ト 精神　名詞, 一般, *, *, *, *, 精神, セイシン, セイシン 的　名詞, 接尾, 形容動詞語幹, *, *, *, 的, テキ, テキ 現象　名詞, 一般, *, *, *, *, 現象, ゲンショウ, ゲンショー を　助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ 暗示　名詞, サ変接続, *, *, *, *, 暗示, アンジ, アンジ し　動詞, 自立, *, *, サ変・スル, 連用形, する, シ, シ て　助詞, 接続助詞, *, *, *, *, て, テ, テ いる　動詞, 非自立, *, *, 一段, 基本形, いる, イル, イル 。　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab fails at identifying the "hyphen" correctly and interprets it as a vowel extend. This is probably due to the fact that hyphens do not exist in Japanese at all. Adding to the fact that かな〜 is a weak interrogative modal final particle, しみ also means stain (a proper noun). Owing to this, the wakachi of the sentence was erroneously done and produced an incorrect tagging. The rest of the sentence is correctly tagged, however.

English Sentence	The tree, as the rule of the physical universe would have it, reflects **light**.
Japanese Sentence	樹木も、物理学の法則によって、**光**を反射している。
MeCab Output	樹木　名詞, 一般, *, *, *, *, 樹木, ジュモク, ジュモク も　助詞, 係助詞, *, *, *, *, も, モ, モ 、　記号, 読点, *, *, *, *, 、, 、, 、 物理　名詞, 一般, *, *, *, *, 物理, ブツリ, ブツリ 学　名詞, 接尾, 一般, *, *, *, 学, ガク, ガク の　助詞, 連体化, *, *, *, *, の, ノ, ノ 法則　名詞, 一般, *, *, *, *, 法則, ホウソク, ホーソク によって　助詞, 格助詞, 連語, *, *, *, によって, ニヨッテ, ニヨッテ 、　記号, 読点, *, *, *, *, 、, 、, 、 **光**　名詞, 一般, *, *, *, *, 光, ヒカリ, ヒカリ を　助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ 反射　名詞, サ変接続, *, *, *, *, 反射, ハンシャ, ハンシャ し　動詞, 自立, *, *, サ変・スル, 連用形, する, シ, シ て　助詞, 接続助詞, *, *, *, *, て, テ, テ いる　動詞, 非自立, *, *, 一段, 基本形, いる, イル, イル 。　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (noun). As there is no okurigana (syllabic endings) it is quite unambiguous in the Japanese version of the sentence. The case particles also help in identifying the word unanimously.

English Sentence	He gave Blanche the **water-color** sketch.
Japanese Sentence	彼は**水彩絵**をブランチェに与えた。
MeCab Output	彼　名詞, 代名詞, 一般, *, *, *, 彼, カレ, カレ は　助詞, 係助詞, *, *, *, *, は, ハ, ワ **水彩**　名詞, 一般, *, *, *, *, 水彩, スイサイ, スイサイ **絵**　名詞, 接尾, 一般, *, *, *, 絵, エ, エ を　助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ ブランチェ　名詞, 一般, *, *, *, *, * に　助詞, 格助詞, 一般, *, *, *, に, ニ, ニ 与え　動詞, 自立, *, *, 一段, 連用形, 与える, アタエ, アタエ た　助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ 。　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (noun-noun). Since there are no syllabic modifiers, they can only be interpreted as two nouns. The adjectival meaning does not even exist in Japanese, as even a chunk like ”water-colored” would be translated as 「水彩で描いた〇」 or ”painted with water-colors”.

English Sentence	You have eyes like a Siberian Husky; underneath the red they are ice **blue**.
Japanese Sentence	あなたの目はシベリアンハスキーのよう、紅色の下は氷のような**青色**だ。
MeCab Output	あなた　名詞, 代名詞, 一般, *, *, *, あなた, アナタ, アナタ の　助詞, 連体化, *, *, *, *, の, ノ, ノ 目　名詞, 一般, *, *, *, *, 目, メ, メ は　助詞, 係助詞, *, *, *, *, は, ハ, ワ シベリアンハスキー　名詞, 固有名詞, 一般, *, *, *, シベリアンハスキー, シベリアンハスキー, シベリアンハスキー の　助詞, 連体化, *, *, *, *, の, ノ, ノ よう　名詞, 非自立, 助動詞語幹, *, *, *, よう, ヨウ, ヨー 、　記号, 読点, *, *, *, *, 、, 、, 、 紅色　名詞, 一般, *, *, *, *, 紅色, コウショク, コーショク の　助詞, 連体化, *, *, *, *, の, ノ, ノ 下　名詞, 一般, *, *, *, *, 下, シタ, シタ は　助詞, 係助詞, *, *, *, *, は, ハ, ワ 氷　名詞, 一般, *, *, *, *, 氷, コオリ, コーリ の　助詞, 連体化, *, *, *, *, の, ノ, ノ よう　名詞, 非自立, 助動詞語幹, *, *, *, よう, ヨウ, ヨー な　助動詞, *, *, *, 特殊・ダ, 体言接続, だ, ナ, ナ **青色**　名詞, 一般, *, *, *, *, 青色, アオイロ, アオイロ だ　助動詞, *, *, *, 特殊・ダ, 基本形, だ, ダ, ダ 。　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors(noun). Colors in Japanese can be adjectives, but as it is combined with ”ice” the potential adjective phrase is turned into a string of nouns.

English Sentence	I liked **that little recitative** by Ilia in Idomeneo.
Japanese Sentence	私はイリアのイドメネーオでのあの**小さな**叙唱が好きだったよ。
MeCab Output	私 名詞, 代名詞, 一般, *, *, *, 私, ワタシ, ワタシ は 助詞, 係助詞, *, *, *, *, は, ハ, ワ イリア 名詞, 一般, *, *, *, *, * の 助詞, 連体化, *, *, *, *, の, ノ, ノ イドメネーオ 名詞, 一般, *, *, *, *, * で 助詞, 格助詞, 一般, *, *, *, で, デ, デ の 助詞, 連体化, *, *, *, *, の, ノ, ノ **あの** 連体詞, *, *, *, *, *, あの, アノ, アノ **小さな** 連体詞, *, *, *, *, *, 小さな, チイサナ, チーサナ **叙** 名詞, サ変接続, *, *, *, *, 叙, ジョ, ジョ **唱** 名詞, 一般, *, *, *, *, * が 助詞, 格助詞, 一般, *, *, *, が, ガ, ガ 好き 名詞, 形容動詞語幹, *, *, *, *, 好き, スキ, スキ だっ 助動詞, *, *, *, 特殊・ダ, 連用タ接続, だ, ダッ, ダッ た 助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ よ 助詞, 終助詞, *, *, *, *, よ, ヨ, ヨ 。 記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (adjective-noun). The only notable part about the tagging is the fact that the determinant that actually becomes part of the nominal adjective in the Japanese sentence. The noun is once again unambiguous due to the nature of sino-japanese compounds.

English Sentence	Here is the face of a patient, **drugged** out of pain and sorrow, drifting on Demerol.
Japanese Sentence	これは苦痛と悲しみを消すために、デメロルを**飲まされた**病人の表情だ。
MeCab Output	これ 名詞, 代名詞, 一般, *, *, *, これ, コレ, コレ は 助詞, 係助詞, *, *, *, *, は, ハ, ワ 苦痛 名詞, 一般, *, *, *, *, 苦痛, クツウ, クツー と 助詞, 並立助詞, *, *, *, *, と, ト, ト 悲しみ 名詞, 一般, *, *, *, *, 悲しみ, カナシミ, カナシミ を 助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ 消す 動詞, 自立, *, *, 五段・サ行, 基本形, 消す, ケス, ケス ため 名詞, 非自立, 副詞可能, *, *, *, ため, タメ, タメ に 助詞, 格助詞, 一般, *, *, *, に, ニ, ニ 、 記号, 読点, *, *, *, *, 、, 、, 、 デメロル 名詞, 一般, *, *, *, *, * を 助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ **飲まさ** 動詞, 自立, *, *, 五段・サ行, 未然形, 飲ます, ノマサ, ノマサ **れ** 動詞, 接尾, *, *, 一段, 連用形, れる, レ, レ **た** 助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ 病人 名詞, 一般, *, *, *, *, 病人, ビョウニン, ビョーニン の 助詞, 連体化, *, *, *, *, の, ノ, ノ

	表情　　　　名詞, 一般, *, *, *, *, 表情, ヒョウジョウ, ヒョージョー だ　　　　　助動詞, *, *, *, 特殊・ダ, 基本形, だ, ダ, ダ 。　　　　　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (verb). Since there is no natural verb or adjective in Japanese directly corresponding to ”drugged”, the counterpart ”was forced to take” had to be used instead to maintain language integrity. This of course circumvented the tagging problem all together. But even if the more unnatural gairaigo (foreign word) 「ドラッグされた」”doraggu sareta” had been used, the fact that the actual verb ”された” “was forced to” is completely removed from the noun base ”ドラッグ” “drug” ensures that it will remain correctly tagged.

English Sentence	That baby was nearly **electrocuted** in the Casino.
Japanese Sentence	あの赤ん坊は危うくカジノで**感電死させられる**ところだった。
MeCab Output	あの　　　　連体詞, *, *, *, *, *, あの, アノ, アノ 赤ん坊　　　名詞, 一般, *, *, *, *, 赤ん坊, アカンボウ, アカンボー は　　　　　助詞, 係助詞, *, *, *, *, は, ハ, ワ 危うく　　　形容詞, 自立, *, *, 形容詞・アウオ段, 連用テ接続, 危うい, アヤウク, アヤウク カジノ　　　名詞, 一般, *, *, *, *, カジノ, カジノ, カジノ で　　　　　助詞, 格助詞, 一般, *, *, *, で, デ, デ **感電**　　　　名詞, サ変接続, *, *, *, *, 感電, カンデン, カンデン **死**　　　　　名詞, 接尾, サ変接続, *, *, *, 死, シ, シ **さ**　　　　　動詞, 自立, *, *, サ変・スル, 未然レル接続, する, サ, サ **せ**　　　　　動詞, 接尾, *, *, 一段, 未然形, せる, セ, セ **られる**　　　動詞, 接尾, *, *, 一段, 基本形, られる, ラレル, ラレル ところ　　　名詞, 非自立, 副詞可能, *, *, *, ところ, トコロ, トコロ だっ　　　　助動詞, *, *, *, 特殊・ダ, 連用タ接続, だ, ダッ, ダッ た　　　　　助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ 。　　　　　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (noun-verb). This ambiguity also seems to be resolved with the help of the sino-japanese. The correlation with the verb is in the okurigana (the ending syllabic characters following the sino-japanese) as always.

English Sentence	Somebody has come into camp with a **stringed** instrument and is singing about Divine Love in Provencal.
Japanese Sentence	誰かが陣営に**弦**を持って入って来て、プロヴァンス語で神の愛を歌い始めた。
MeCab Output	誰　名詞, 代名詞, 一般, *, *, *, 誰, ダレ, ダレ か　助詞, 副助詞／並立助詞／終助詞, *, *, *, *, か, カ, カ が　助詞, 格助詞, 一般, *, *, *, が, ガ, ガ 陣営　名詞, 一般, *, *, *, *, 陣営, ジンエイ, ジンエイ に　助詞, 格助詞, 一般, *, *, *, に, ニ, ニ **弦**　名詞, 一般, *, *, *, *, 弦, ツル, ツル を　助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ 持っ　動詞, 自立, *, *, 五段・タ行, 連用タ接続, 持つ, モッ, モッ て　助詞, 接続助詞, *, *, *, *, て, テ, テ 入っ　動詞, 自立, *, *, 五段・ラ行, 連用タ接続, 入る, ハイッ, ハイッ て　助詞, 接続助詞, *, *, *, *, て, テ, テ 来　動詞, 非自立, *, *, カ変・来ル, 連用形, 来る, キ, キ て　助詞, 接続助詞, *, *, *, *, て, テ, テ 、　記号, 読点, *, *, *, *, 、, 、, 、 プロヴァンス　名詞, 固有名詞, 地域, 一般, *, *, プロヴァンス, プロヴァンス, プロバンス 語　名詞, 接尾, 一般, *, *, *, 語, ゴ, ゴ で　助詞, 格助詞, 一般, *, *, *, で, デ, デ 神　名詞, 一般, *, *, *, *, 神, カミ, カミ の　助詞, 連体化, *, *, *, *, の, ノ, ノ 愛　名詞, 一般, *, *, *, *, 愛, アイ, アイ を　助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ 歌い　動詞, 自立, *, *, 五段・ワ行促音便, 連用形, 歌う, ウタイ, ウタイ 始め　動詞, 非自立, *, *, 一段, 連用形, 始める, ハジメ, ハジメ た　助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ 。　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (noun). This example is interesting due to the fact that not only the conflicting word was resolved but also that words surrounding the word are actually condensed into one kanji 「弦」which can mean everything from string to a piece of music performed by a stringed instrument.

English Sentence	Our own barbarians **vs** theirs now!
Japanese Sentence	今度は、我々の野蛮人**に対して**敵の野蛮人！
MeCab Output	今度 名詞, 副詞可能, *, *, *, *, 今度, コンド, コンド は 助詞, 係助詞, *, *, *, *, は, ハ, ワ 、 記号, 読点, *, *, *, *, 、, 、, 、 我々 名詞, 代名詞, 一般, *, *, *, 我々, ワレワレ, ワレワレ の 助詞, 連体化, *, *, *, *, の, ノ, ノ 野蛮 名詞, 一般, *, *, *, *, 野蛮, ヤバン, ヤバン 人 名詞, 接尾, 一般, *, *, *, 人, ジン, ジン **に対して** 助詞, 格助詞, 連語, *, *, *, に対して, ニタイシテ, ニタイシテ 敵 名詞, 一般, *, *, *, *, 敵, テキ, テキ の 助詞, 連体化, *, *, *, *, の, ノ, ノ 野蛮 名詞, 一般, *, *, *, *, 野蛮, ヤバン, ヤバン 人 名詞, 接尾, 一般, *, *, *, 人, ジン, ジン ！ 記号, 一般, *, *, *, *, ！, ！, ！ EOS
Comment	MeCab tags the sentence without errors (particle compound). This sentence does not translate well into Japanese at all. For one, the word ”barbarians” bears very different connotations, meaning either the northern ”wild people”; the Ainu from the Hokkaido region or equally, us westerners. The concept of ”our barbarians” simply does not really exist. The initial result of the tag of ”particle”, as the POS is four characters long, also baffled us quite a bit, but after taking the ”compound” part of the tag into account, it finally made a bit more sense.

English Sentence	Siegfried **sniffles** as his eyes refocus on the Autobahn.
Japanese Sentence	シーグフリートはアウトバーンに再び視線を向けた瞬間、**鼻をすすった**。
MeCab Output	シーグフリート 名詞, 一般, *, *, *, *, * は 助詞, 係助詞, *, *, *, *, は, ハ, ワ アウトバーン 名詞, 一般, *, *, *, *, アウトバーン, アウトバーン, アウトバーン に 助詞, 格助詞, 一般, *, *, *, に, ニ, ニ 再び 副詞, 助詞類接続, *, *, *, *, 再び, フタタビ, フタタビ 視線 名詞, 一般, *, *, *, *, 視線, シセン, シセン を 助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ 向け 動詞, 自立, *, *, 一段, 連用形, 向ける, ムケ, ムケ た 助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ 瞬間 名詞, 副詞可能, *, *, *, *, 瞬間, シュンカン, シュンカン 、 記号, 読点, *, *, *, *, 、, 、, 、 **鼻** 名詞, 一般, *, *, *, *, 鼻, ハナ, ハナ **を** 助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ **すすっ** 動詞, 自立, *, *, 五段・ラ行, 連用タ接続, すする, ススッ, ススッ **た** 助動詞, *, *, *, 特殊・タ, 基本形, た, タ, タ

	。　　　　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (noun-particle-verb). The Japanese version of ”sniffles” is an idiom encompassing a sino-japanese character. As such, no ambiguity arises.

English Sentence	Caprice **worries** about her husband.
Japanese Sentence	カプリスは夫の**心配をしている**。
MeCab Output	カプリス　　名詞, 一般, *, *, *, *, * は　　助詞, 係助詞, *, *, *, *, は, ハ, ワ 夫　　名詞, 一般, *, *, *, *, 夫, オット, オット の　　助詞, 連体化, *, *, *, *, の, ノ, ノ **心配**　　名詞, サ変接続, *, *, *, *, 心配, シンパイ, シンパイ **を**　　助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ **し**　　動詞, 自立, *, *, サ変・スル, 連用形, する, シ, シ **て**　　助詞, 接続助詞, *, *, *, *, て, テ, テ **いる**　　動詞, 非自立, *, *, 一段, 基本形, いる, イル, イル 。　　記号, 句点, *, *, *, *, 。, 。, 。 EOS
Comment	MeCab tags the sentence without errors (noun-particle-verb). Yet again, the ambiguous English word is replaced by a noun-verb combination starting with an ideogram.

Some conclusions can be drawn from these data. Firstly, the accuracy for MeCab is about the same as the Stanford tagger; generally a bit over 90%[9][14]. The features most common to cause problems for the Stanford tagger with English sentences does not, however, seem to pose any problems to MeCab. This tells us that the nature of the error margin must be different for the two languages. The error might be due to bad correlation between the two languages, or even differences between the two taggers used. But both taggers implemented a maximum entropy algorithm. And looking at the translations and analyzing just the part that is problematic in the English texts, we can see that the Japanese stems still are just as ambiguous as the English ones. The difference is that the Japanese grammar uses okurigana as markers to help the tagger identify words. It also chunks words more than the Stanford one. for example splitting sino-japanese adjectives into a noun part and the adjectival modifier.

We thought that one reason for the high success rate of the test data we provided to MeCab might have been because of the liberal use of kanji, so we ran a second test, with all kanji rewritten as kana. The results can be seen in the appendix, the erroneously tagged parts have been marked with cursive font style. As one can see the error rate is over 50% per sentence when not using kanji. We could therefore argue that the error rate in Japanese in tagging effectively lies with wakachi problems, as hinted at the start while English tagging errors are more in the domain of actual word ambiguity. One has to point out, however, that some of the errors produced in the second test is due to the fact that kanji compounds should not be written as kana. A second test could be made, translating the kango to wago to see if it has any additional effect on the output.

The second test suggests that the Japanese tagger is very sensitive to misspelled words. This would not be inconceivable as there are no real delimiters in the text, so when one word has been tokenized incorrectly, that error quickly propagates. This can be compared to the English version of test 6 where one erroneous tag quickly propagates its error to the other words. This reason for this juxtaposition becomes apparent when we take the Japanese (case) markers into account. Since each token is a smaller semantic chunk in Japanese, the possible tags for each token decreases. And as such, errors do not propagate as easily.

4.1.6 Japanese Tounge Twisters

For some further testing with the kanji / wago aspect we tried running some well known wago tongue twisters through the program. All results are in the appendix.

裏庭 には 二羽、 庭 には 二羽、鶏 あり
Uraniwa niwa niwa, niwa niwa niwa, niwatori ari.
There are two chickens in the backyard, and two in the garden.

李 も 桃 も 桃 の うち、桃 も 李 も 桃 の うち
Sumomo mo momo mo momo no uchi, momo mo sumomo mo momo no uchi.
Both plums and peaches are a kind of peach. Both peaches and plums are a kind of peach.

瓜売り が 瓜 売りに 来て 瓜 売れず 売り 売り かえる 瓜売り の 声
Uriuri ga uri urini kite uri urezu uri uri kaeru uriuri no koe.
A melon vender came to sell melons, but no melon was sold; O, the voice of melon vender who is back, selling melons.

All three kanji saturated sentences are, as expected, tagged correctly.
Interestingly enough, the kanjiless sentences were either complete disasters, or did very well.
The first tonguetwister, with the wakachi used by the tagger essentially reads:
"An alligator-haniwa in the behind, to alligator-haniwa, there is chicken". (A haniwa is an ancient Japanese clay figurine) While the other two (I must say, a bit to our surprise) were completely correct.

It basically boils down to word length and sentence distribution; short length is more common, so a sentence with many short words have a large probability to be tagged correctly when the kanji is not present.

The sentence distribution becomes important since if an unusual word is placed early in the sentence, and a substring of that word happens to be a common word, the wakachi will almost always do the split wrong and propagate that error through the entire sentence.
The tagger becomes very unstable without the kanji though, as even one erroneous tag at the start of the sentence propagates directly to the end, as can be seen in the first example.

5 Conclusions

As pointed out early in this report, the intended goal was to "*determine whether the errors follow a pattern within a language, and if the same errors appear consistently between different languages* "

As could clearly be seen in section 4.1.2 and 4.1.3, errors within a language does indeed follow a pattern which is most obviously visible in section 4.1.2, which identifies the most common failure pattern as misclassifications of words of the types adjective and noun.
More patterns were identified as many of the analyzed sentences were discovered to most likely have been misclassified due to the improbability of the sequence, it can be concluded that the inherent weakness in maximum entropy taggers is their tendency to systematically fail on sentences with rare sequences of word types. Even though taggers like the Stanford POST takes into consideration not only the two tags preceding the current word, but also one word superseding the current word to be evaluated. This makes the answer to the first question a resounding yes.

Sections 4.1.4 and 4.1.5 later demonstrate that the identified patterns are not consistent between languages that are not grammatically compatible as the translated sentences, although they should have the same adjective / noun issue, were almost all tagged correctly. Instead we managed to find a different pattern of failure which was Japanese taggers weakness to sentences featuring a lot of kana. The reason for this was identified as difficulties in chunking sentences in Japanese given how kana, depending on the context, can mean a lot of different things.

The second goal which was also pointed out early was to draw a conclusion as to what could be done to improve current tagging technices.

The idea we arrived at, seeing how the earlier identified patterns of failure were mostly due to how strictly statisticall the taggers we used are is to improve accuracy by providing the algorithms with an aiding lookup table. The algorithm would then very roughly work as described with the following pseudo code

```
W← increase word context with one word
IF the lookup table contains a unique tag solution for W THEN
        tag W with the unique tag(s)
ELSE IF the lookup table contains no entry for W THEN
        use the old statisticall methods to tag W
ELSE
        recurse to start of algorithm
END IF
```

We feel this will work mostly because it couldn't possibly make the results worse; the algorithm starts out with just one word to tag, if that has a unique solution then most surely it will be correct or the training data must be horribly wrong. As the sentence grows larger, the only cases where the lookup table will contain unique solutions will be for figures of speech, something which is most surely correctly tagged in the training data and for any sentencs which are shown to have no unique solutions, the context will expand to the point where this is either the case, or the tagger will simply resort to using the old statisticall methods for evaluating the tags.

6 Bibliography

[1] al., G. G. (n.d.). *Dialogues with colorful personalities of early ai.* Retrieved March 7, 2010, from http://www.stanford.edu/group/SHR/4-2/text/dialogues.html

[2] *American National Corpus.* (n.d.). Retrieved March 11, 2010, from Open ANC: http://www.americannationalcorpus.org/OANC/

[3] *GNU General Public License.* (u.d.). Hämtat från http://www.gnu.org/copyleft/gpl.html den 11 March 2010

[4] Kudo, T. (n.d.). *MeCab: Yet Another Part-of-Speech and Morphological Analyzer.* Retrieved March 11, 2010, from http://mecab.sourceforge.net/

[5] Megyesi, B. (2002). *Data-Driven Syntactic Analysis Methods and Applications for Swedish.* Stockholm.

[6] Sigurd, B. (1994). *Computerized Grammars for Analysis and Machine Translation.* Lund: Lund University Press.

[7] *Stanford Log-linear Part-Of-Speech Tagger.* (n.d.). Retrieved Match 11, 2010, from The Stanford Natural Language Processing Group: http://nlp.stanford.edu/software/tagger.shtml

[8] *Stockholm Umeå Corpus.* (u.d.). Hämtat från http://www.ling.su.se/staff/sofia/suc/suc.html den 11 March 2010

[9] (n.d.). Retrieved 05 01, 2010, from http://pythonlife.seesaa.net/article/136098000.html

[10] *Morphological Tags.* (n.d.). Retrieved 04 28, 2010, from www2.lingsoft.fi/doc/swecg/intro/mtags.html

[11] *Parole Tag set.* (n.d.). Retrieved 04 28, 2010, from www.lsi.upc.edu/~nlp/SVMTool/parole.html

[12] Pinker, S. (1994). *The Language Instinct.*

[13] *Språkbanken.* (n.d.). Retrieved 04 28, 2010, from http://spraakbanken.gu.se/lb/parole/

[14] Villodre, L. M. (n.d.). *Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees.* Retrieved 05 01, 2010, from http://www.tesisenxarxa.net/TESIS_UPC/AVAILABLE/TDX-0723109-124136//TLMV1de2.pdf

[15] James Allan, H. R. (u.d.). *Using part-of-speech patterns to reduce query ambiguity.* Retrieved 05 01, 2010, from http://portal.acm.org/citation.cfm?id=564430&dl=GUIDE&coll=GUIDE&CFID=86938427&CFTOKEN=16194787

[16] Rohit J. Kate, Y. W. (n.d.). *Learning to Transform Natural to Formal Languages.* Retrieved 05 02, 2010, from http://userweb.cs.utexas.edu/users/ml/papers/transform-aaai-05.pdf

Appendix A

Pen Treebank tag set

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NP	Proper noun, singular
15.	NPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PP	Personal pronoun
19.	PP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Appendix B

Japanese terminology primer

Kanji - ideomatic written characters, also referred to as sino-japanese, can be standalone or appear in compounds. The meaning of a compound can have a very different meaning from what it constituents mean, and can have different grammatical roles depending on context and **okurigana**.

Okurigana - **kana** used after **kanji** and **kanji compounds** to indicate morphology and other grammatical traits.

Kana - the two syllabic written language systems used in Japan. Hiragana is used for native purposes, katakana is often used in technology / computer science settings and for foreign words.

Wakachi - a type of tokenizing. Since Japanese does not use spacing in text, but relies on the relationships between **kanji** and **kana**, wakachi poses a much greater threat to tagger accuracy in Japanese than it would in English.

Gairaigo - imported foreign terms written in **kana.**

Kango - Imported Chinese word versions, often written as kanji compounds with okurigana.

Wago - Native Japanese word versions, often written as singleton kanji with okurigana.

Appendix C

Results of translated sentences rewritten with kana

```
じゃ        接続詞,*,*,*,*,*,じゃ,ジャ,ジャ
あく        名詞,一般,*,*,*,*,あく,アク,アク
な          助動詞,*,*,*,特殊・ダ,体言接続,だ,ナ,ナ
おんな      名詞,一般,*,*,*,*,おんな,オンナ,オンナ
が          助詞,格助詞,一般,*,*,*,が,ガ,ガ
よ          助詞,終助詞,*,*,*,*,よ,ヨ,ヨ
なき        形容詞,自立,*,*,形容詞・アウオ段,体言接続,ない,ナキ,ナキ
する        動詞,自立,*,*,サ変・スル,基本形,する,スル,スル
あかん      感動詞,*,*,*,*,*,あかん,アカン,アカン
ぼう        名詞,一般,*,*,*,*,ぼう,ボウ,ボウ
を          助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
ほ          動詞,自立,*,*,五段・ラ行,体言接続特殊2,ほる,ホ,ホ
に          助詞,格助詞,一般,*,*,*,に,ニ,ニ
ゅうびんにはいった          名詞,一般,*,*,*,*,*
ワイン      名詞,一般,*,*,*,*,ワイン,ワイン,ワイン
で          助詞,格助詞,一般,*,*,*,で,デ,デ
あやし      動詞,自立,*,*,五段・サ行,連用形,あやす,アヤシ,アヤシ
て          助詞,接続助詞,*,*,*,*,て,テ,テ
いる        動詞,非自立,*,*,一段,基本形,いる,イル,イル
。          記号,句点,*,*,*,*,。,。,。
EOS

この        連体詞,*,*,*,*,*,この,コノ,コノ
か          助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ
なー        助詞,終助詞,*,*,*,*,なー,ナー,ナー
しみ        名詞,一般,*,*,*,*,しみ,シミ,シミ
そのもの    名詞,一般,*,*,*,*,そのもの,ソノモノ,ソノモノ
が          助詞,格助詞,一般,*,*,*,が,ガ,ガ
すでに      副詞,一般,*,*,*,*,すでに,スデニ,スデニ
びょう      名詞,一般,*,*,*,*,びょう,ビョウ,ビョー
に          助詞,格助詞,一般,*,*,*,に,ニ,ニ
ん          名詞,非自立,一般,*,*,*,ん,ン,ン
の          助詞,連体化,*,*,*,*,の,ノ,ノ
にく        形容詞,自立,*,*,形容詞・アウオ段,ガル接続,にくい,ニク,ニク
たい        動詞,自立,*,*,五段・カ行イ音便,連用タ接続,たく,タイ,タイ
て          助詞,接続助詞,*,*,*,*,て,テ,テ
き          動詞,非自立,*,*,カ変・クル,連用形,くる,キ,キ
かんかく    名詞,サ変接続,*,*,*,*,かんかく,カンカク,カンカク
とせい      名詞,一般,*,*,*,*,とせい,トセイ,トセイ
しん        動詞,自立,*,*,五段・マ行,連用タ接続,しむ,シン,シン
て          助詞,接続助詞,*,*,*,*,て,テ,テ
き          動詞,非自立,*,*,カ変・クル,連用形,くる,キ,キ
げん        名詞,一般,*,*,*,*,げん,ゲン,ゲン
しょう      名詞,一般,*,*,*,*,しょう,ショウ,ショー
を          助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
あんじ      動詞,自立,*,*,一段,連用形,あんじる,アンジ,アンジ
し          動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
て          助詞,接続助詞,*,*,*,*,て,テ,テ
いる        動詞,非自立,*,*,一段,基本形,いる,イル,イル
```

```
。          記号,句点,*,*,*,*,。,。,。
EOS
じ          助動詞,*,*,*,不変化型,基本形,じ,ジ,ジ
ゅもくも    名詞,固有名詞,組織,*,*,*,*
、          記号,読点,*,*,*,*,、,、,、
ぶつ        名詞,一般,*,*,*,*,ぶつ,ブツ,ブツ
り          助動詞,*,*,*,文語・リ,基本形,り,リ,リ
がく        名詞,一般,*,*,*,*,がく,ガク,ガク
の          助詞,連体化,*,*,*,*,の,ノ,ノ
ほう        名詞,非自立,一般,*,*,*,ほう,ホウ,ホー
そく        名詞,一般,*,*,*,*,そく,ソク,ソク
によって    助詞,格助詞,連語,*,*,*,によって,ニヨッテ,ニヨッテ
、          記号,読点,*,*,*,*,、,、,、
ひかり      名詞,固有名詞,一般,*,*,*,ひかり,ヒカリ,ヒカリ
を          助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
はん        動詞,自立,*,*,五段・マ行,連用タ接続,はむ,ハン,ハン
しゃ        動詞,接尾,*,*,五段・サ行,仮定縮約1,す,シャ,シャ
し          動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
て          助詞,接続助詞,*,*,*,*,て,テ,テ
いる        動詞,非自立,*,*,一段,基本形,いる,イル,イル
。          記号,句点,*,*,*,*,。,。,。

EOS
かれ        動詞,自立,*,*,一段,連用形,かれる,カレ,カレ
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
すい        動詞,自立,*,*,五段・ワ行促音便,連用形,すう,スイ,スイ
さ          名詞,接尾,特殊,*,*,*,さ,サ,サ
いえ        動詞,自立,*,*,一段,連用形,いえる,イエ,イエ
を          助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
ブランチェ  名詞,一般,*,*,*,*,*
に          助詞,格助詞,一般,*,*,*,に,ニ,ニ
あたえ      動詞,自立,*,*,一段,連用形,あたえる,アタエ,アタエ
た          助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。          記号,句点,*,*,*,*,。,。,。

EOS
あなた      名詞,代名詞,一般,*,*,*,あなた,アナタ,アナタ
の          助詞,連体化,*,*,*,*,の,ノ,ノ
め          名詞,一般,*,*,*,*,め,メ,メ
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
シベリアンハスキー          名詞,固有名詞,一般,*,*,*,シベリアンハスキー,シベリアンハスキー,シベリアン
ハスキー
の          助詞,連体化,*,*,*,*,の,ノ,ノ
よう        名詞,非自立,助動詞語幹,*,*,*,よう,ヨウ,ヨー
、          記号,読点,*,*,*,*,、,、,、
べ          助詞,終助詞,*,*,*,*,べ,べ,べ
に          助詞,格助詞,一般,*,*,*,に,ニ,ニ
いろ        動詞,自立,*,*,一段,命令ro,いる,イロ,イロ
の          助詞,連体化,*,*,*,*,の,ノ,ノ
し          動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た          助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
こおり      動詞,自立,*,*,五段・ラ行,連用形,こおる,コオリ,コーリ
の          助詞,連体化,*,*,*,*,の,ノ,ノ
```

```
よう        名詞,非自立,助動詞語幹,*,*,*,よう,ヨウ,ヨー
な          助動詞,*,*,*,特殊・ダ,体言接続,だ,ナ,ナ
あおい      形容詞,自立,*,*,形容詞・アウオ段,基本形,あおい,アオイ,アオイ
ろ          名詞,一般,*,*,*,*,ろ,ロ,ロ
だ          助動詞,*,*,*,特殊・ダ,基本形,だ,ダ,ダ
。          記号,句点,*,*,*,*,。,。,。
EOS

わたし      名詞,代名詞,一般,*,*,*,わたし,ワタシ,ワタシ
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
イリア      名詞,一般,*,*,*,*,*
の          助詞,連体化,*,*,*,*,の,ノ,ノ
イドメネーオ名詞,一般,*,*,*,*,*
で          助詞,格助詞,一般,*,*,*,で,デ,デ
の          助詞,連体化,*,*,*,*,の,ノ,ノ
あの        連体詞,*,*,*,*,*,あの,アノ,アノ
ちいさな    連体詞,*,*,*,*,*,ちいさな,チイサナ,チーサナ
じょしょ    動詞,自立,*,*,サ変・－スル,未然ウ接続,じょする,ジョショ,ジョショ
う          助動詞,*,*,*,不変化型,基本形,う,ウ,ウ
が          助詞,接続助詞,*,*,*,*,が,ガ,ガ
すき        名詞,一般,*,*,*,*,すき,スキ,スキ
だっ        助動詞,*,*,*,特殊・ダ,連用タ接続,だ,ダッ,ダッ
た          助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
よ          助詞,終助詞,*,*,*,*,よ,ヨ,ヨ
。          記号,句点,*,*,*,*,。,。,。
EOS

これ        名詞,代名詞,一般,*,*,*,これ,コレ,コレ
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
くつ        名詞,一般,*,*,*,*,くつ,クツ,クツ
う          助動詞,*,*,*,不変化型,基本形,う,ウ,ウ
と          助詞,格助詞,引用,*,*,*,と,ト,ト
かなしみ    名詞,一般,*,*,*,*,かなしみ,カナシミ,カナシミ
を          助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
けす        動詞,自立,*,*,五段・サ行,基本形,けす,ケス,ケス
ため        名詞,非自立,副詞可能,*,*,*,ため,タメ,タメ
に          助詞,格助詞,一般,*,*,*,に,ニ,ニ
、          記号,読点,*,*,*,*,、,、,、
デメロル    名詞,一般,*,*,*,*,*
を          助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
のまさ      動詞,自立,*,*,五段・サ行,未然形,のます,ノマサ,ノマサ
れ          動詞,接尾,*,*,一段,連用形,れる,レ,レ
たび        名詞,非自立,副詞可能,*,*,*,たび,タビ,タビ
ょうにんのひょうじょうだ    名詞,一般,*,*,*,*,*
。          記号,句点,*,*,*,*,。,。,。
EOS

あの        フィラー,*,*,*,*,*,あの,アノ,アノ
あかん      感動詞,*,*,*,*,*,あかん,アカン,アカン
ぼう        名詞,一般,*,*,*,*,ぼう,ボウ,ボウ
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
あやうく    形容詞,自立,*,*,形容詞・アウオ段,連用テ接続,あやうい,アヤウク,アヤウク
カジノ      名詞,一般,*,*,*,*,カジノ,カジノ,カジノ
で          助詞,格助詞,一般,*,*,*,で,デ,デ
```

```
かん	名詞,一般,*,*,*,*,かん,カン,カン
でん	副詞,助詞類接続,*,*,*,*,でん,デン,デン
しさ	動詞,自立,*,*,五段・サ行,未然形,しす,シサ,シサ
せ	動詞,接尾,*,*,一段,未然形,せる,セ,セ
られる	動詞,接尾,*,*,一段,基本形,られる,ラレル,ラレル
ところ	名詞,非自立,副詞可能,*,*,*,ところ,トコロ,トコロ
だっ	助動詞,*,*,*,特殊・ダ,連用タ接続,だ,ダッ,ダッ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。
EOS

だれ	名詞,代名詞,一般,*,*,*,だれ,ダレ,ダレ
か	助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
ぐんえいにげんをもってはいってきて	名詞,一般,*,*,*,*,*
、	記号,読点,*,*,*,*,、,、,、
プロヴァンス名詞,固有名詞,地域,一般,*,*,プロヴァンス,プロヴァンス,プロバンス
ご	接頭詞,名詞接続,*,*,*,*,ご,ゴ,ゴ
でか	形容詞,自立,*,*,形容詞・アウオ段,ガル接続,でかい,デカ,デカ
み	名詞,接尾,特殊,*,*,*,み,ミ,ミ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
あい	名詞,一般,*,*,*,*,あい,アイ,アイ
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
うたい	動詞,自立,*,*,五段・ワ行促音便,連用形,うたう,ウタイ,ウタイ
はじめ	動詞,非自立,*,*,一段,連用形,はじめる,ハジメ,ハジメ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。
EOS

こんど	名詞,副詞可能,*,*,*,*,こんど,コンド,コンド
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
、	記号,読点,*,*,*,*,、,、,、
われわれ	名詞,代名詞,一般,*,*,*,われわれ,ワレワレ,ワレワレ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
や	助詞,並立助詞,*,*,*,*,や,ヤ,ヤ
ばんじん	名詞,一般,*,*,*,*,ばんじん,バンジン,バンジン
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
たいし	動詞,自立,*,*,五段・サ行,連用形,たいす,タイシ,タイシ
て	助詞,接続助詞,*,*,*,*,て,テ,テ
てき	動詞,非自立,*,*,五段・カ行促音便,連用形,てく,テキ,テキ
の	名詞,非自立,一般,*,*,*,の,ノ,ノ
や	助詞,並立助詞,*,*,*,*,や,ヤ,ヤ
ばんじん	名詞,一般,*,*,*,*,ばんじん,バンジン,バンジン
！	記号,一般,*,*,*,*,！,！,！
EOS

シーグフリード		名詞,固有名詞,人名,名,*,*,シーグフリード,シーグフリード,シーグフリード
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
アウトバーン		名詞,一般,*,*,*,*,アウトバーン,アウトバーン,アウトバーン
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
ふた	名詞,一般,*,*,*,*,ふた,フタ,フタ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
びしせんをむけたしゅんかん	名詞,一般,*,*,*,*,*
、	記号,読点,*,*,*,*,、,、,、
```

```
はな      名詞,固有名詞,人名,名,*,*,はな,ハナ,ハナ
を        助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
すすっ    動詞,自立,*,*,五段・ラ行,連用タ接続,すする,ススッ,ススッ
た        助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。        記号,句点,*,*,*,*,。,。,。
EOS

カプリス  名詞,一般,*,*,*,*,*
は        助詞,係助詞,*,*,*,*,は,ハ,ワ
おっ      動詞,自立,*,*,五段・ラ行,連用タ接続,おる,オッ,オッ
と        助詞,格助詞,引用,*,*,*,と,ト,ト
の        助詞,連体化,*,*,*,*,の,ノ,ノ
しんぱい  名詞,サ変接続,*,*,*,*,しんぱい,シンパイ,シンパイ
を        助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
し        動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
て        助詞,接続助詞,*,*,*,*,て,テ,テ
いる      動詞,非自立,*,*,一段,基本形,いる,イル,イル
。        記号,句点,*,*,*,*,。,。,。
EOS
```

Appendix D

Japanese tounge twisters

```
裏庭      名詞,一般,*,*,*,*,裏庭,ウラニワ,ウラニワ
に        助詞,格助詞,一般,*,*,*,に,ニ,ニ
は        助詞,係助詞,*,*,*,*,は,ハ,ワ
二        名詞,数,*,*,*,*,二,ニ,ニ
羽        名詞,接尾,助数詞,*,*,*,羽,ワ,ワ
、        記号,読点,*,*,*,*,、,、,、
庭        名詞,一般,*,*,*,*,庭,ニワ,ニワ
に        助詞,格助詞,一般,*,*,*,に,ニ,ニ
は        助詞,係助詞,*,*,*,*,は,ハ,ワ
二        名詞,数,*,*,*,*,二,ニ,ニ
羽        名詞,接尾,助数詞,*,*,*,羽,ワ,ワ
、        記号,読点,*,*,*,*,、,、,、
鶏        名詞,一般,*,*,*,*,鶏,ニワトリ,ニワトリ
あり      動詞,自立,*,*,ラ変,基本形,あり,アリ,アリ
。        記号,句点,*,*,*,*,。,。,。
EOS

うら      動詞,自立,*,*,五段・ラ行,未然形,うる,ウラ,ウラ
に        助詞,格助詞,一般,*,*,*,に,ニ,ニ
わに      名詞,一般,*,*,*,*,わに,ワニ,ワニ
はにわ    名詞,一般,*,*,*,*,はにわ,ハニワ,ハニワ
、        記号,読点,*,*,*,*,、,、,、
に        助詞,格助詞,一般,*,*,*,に,ニ,ニ
わに      名詞,一般,*,*,*,*,わに,ワニ,ワニ
はにわ    名詞,一般,*,*,*,*,はにわ,ハニワ,ハニワ
、        記号,読点,*,*,*,*,、,、,、
にわとり  名詞,一般,*,*,*,*,にわとり,ニワトリ,ニワトリ
あり      動詞,自立,*,*,ラ変,基本形,あり,アリ,アリ
。        記号,句点,*,*,*,*,。,。,。
EOS

李        名詞,固有名詞,人名,姓,*,*,李,リ,リ
も        助詞,係助詞,*,*,*,*,も,モ,モ
桃        名詞,一般,*,*,*,*,桃,モモ,モモ
も        助詞,係助詞,*,*,*,*,も,モ,モ
桃        名詞,一般,*,*,*,*,桃,モモ,モモ
の        助詞,連体化,*,*,*,*,の,ノ,ノ
うち      名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
、        記号,読点,*,*,*,*,、,、,、
桃        名詞,一般,*,*,*,*,桃,モモ,モモ
も        助詞,係助詞,*,*,*,*,も,モ,モ
李        名詞,固有名詞,人名,姓,*,*,李,リ,リ
も        助詞,係助詞,*,*,*,*,も,モ,モ
桃        名詞,一般,*,*,*,*,桃,モモ,モモ
の        助詞,連体化,*,*,*,*,の,ノ,ノ
うち      名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
。        記号,句点,*,*,*,*,。,。,。
EOS
```

```
すもも	名詞,一般,*,*,*,*,すもも,スモモ,スモモ
も	助詞,係助詞,*,*,*,*,も,モ,モ
もも	名詞,一般,*,*,*,*,もも,モモ,モモ
も	助詞,係助詞,*,*,*,*,も,モ,モ
もも	名詞,一般,*,*,*,*,もも,モモ,モモ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
うち	名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
、	記号,読点,*,*,*,*,、,、,、
もも	名詞,一般,*,*,*,*,もも,モモ,モモ
も	助詞,係助詞,*,*,*,*,も,モ,モ
すもも	名詞,一般,*,*,*,*,すもも,スモモ,スモモ
も	助詞,係助詞,*,*,*,*,も,モ,モ
もも	名詞,一般,*,*,*,*,もも,モモ,モモ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
うち	名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
。	記号,句点,*,*,*,*,。,。,。
EOS

瓜	名詞,一般,*,*,*,*,瓜,ウリ,ウリ
売り	動詞,自立,*,*,五段・ラ行,連用形,売る,ウリ,ウリ
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
瓜	名詞,一般,*,*,*,*,瓜,ウリ,ウリ
売り	動詞,自立,*,*,五段・ラ行,連用形,売る,ウリ,ウリ
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
来	動詞,自立,*,*,カ変・来ル,連用形,来る,キ,キ
て	助詞,接続助詞,*,*,*,*,て,テ,テ
瓜	名詞,一般,*,*,*,*,瓜,ウリ,ウリ
売れ	動詞,自立,*,*,一段,未然形,売れる,ウレ,ウレ
ず	助動詞,*,*,*,特殊・ヌ,連用ニ接続,ぬ,ズ,ズ
売り	動詞,自立,*,*,五段・ラ行,連用形,売る,ウリ,ウリ
売り	動詞,自立,*,*,五段・ラ行,連用形,売る,ウリ,ウリ
かえる	動詞,自立,*,*,一段,基本形,かえる,カエル,カエル
瓜	名詞,一般,*,*,*,*,瓜,ウリ,ウリ
売り	動詞,自立,*,*,五段・ラ行,連用形,売る,ウリ,ウリ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
声	名詞,一般,*,*,*,*,声,コエ,コエ
。	記号,句点,*,*,*,*,。,。,。
EOS

うり	名詞,一般,*,*,*,*,うり,ウリ,ウリ
うり	名詞,一般,*,*,*,*,うり,ウリ,ウリ
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
うり	名詞,一般,*,*,*,*,うり,ウリ,ウリ
うり	名詞,一般,*,*,*,*,うり,ウリ,ウリ
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
き	動詞,自立,*,*,カ変・クル,連用形,くる,キ,キ
て	助詞,接続助詞,*,*,*,*,て,テ,テ
うり	名詞,一般,*,*,*,*,うり,ウリ,ウリ
うれ	動詞,自立,*,*,一段,未然形,うれる,ウレ,ウレ
ず	助動詞,*,*,*,特殊・ヌ,連用ニ接続,ぬ,ズ,ズ
うり	名詞,一般,*,*,*,*,うり,ウリ,ウリ
うり	動詞,自立,*,*,五段・ラ行,連用形,うる,ウリ,ウリ
かえる	動詞,自立,*,*,一段,基本形,かえる,カエル,カエル
```

```
うり        名詞,一般,*,*,*,*,うり,ウリ,ウリ
うり        名詞,一般,*,*,*,*,うり,ウリ,ウリ
の          助詞,格助詞,一般,*,*,*,の,ノ,ノ
こえ        動詞,自立,*,*,五段・ワ行促音便,命令e,こう,コエ,コエ
。          記号,句点,*,*,*,*,。,。,。
EOS
```

www.kth.se