Labeling Parts of Speech Using Untrained Annotators on Mechanical Turk

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Jacob Emil Mainzer

Graduate Program in Computer Science and Engineering

The Ohio State University

2011

Master's Examination Committee:

Professor Eric Fosler-Lussier, Advisor

Professor Mikhail Belkin

Abstract

Supervised learning algorithms often require large amounts of labeled data. Creating this data can be time consuming and expensive. Recent work has used untrained annotators on Mechanical Turk to quickly and cheaply create data for NLP tasks, such as word sense disambiguation, word similarity, machine translation, and PP attachment. In this experiment, we test whether untrained annotators can accurately perform the task of POS tagging. We design a Java Applet, called the Interactive Tagging Guide (ITG) to assist untrained annotators in accurately and quickly POS tagging words using the Penn Treebank tagset. We test this Applet on a small corpus using Mechanical Turk, an online marketplace where users earn small payments for the completion of short tasks. Our results demonstrate that, given the proper assistance, untrained annotators are able to tag parts of speech with approximately 90% accuracy. Furthermore, we analyze the performance of expert annotators using the ITG and discover nearly identical levels of performance as compared to the untrained annotators.

Vita

2009.....B.S. Physics, University of Rochester

September 2009 – August 2010.....Distinguished University Fellowship, The Ohio State University

September 2010 – June 2011 .....Graduate Teaching Assistant, The Ohio State University

Fields of Study

Major Field: Computer Science and Engineering

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The advances in automated part of speech tagging can be credited to the availability of large annotated corpora, such as the Penn Treebank, from which linguistic information can be extracted [1]. However, creating large, POS annotated corpora can be time-consuming and expensive. Annotation must be performed by experts, often with graduate training in linguistics and dozens of hours of training. For example, annotators for the Penn Treebank received training 15 hours a week for almost a month [2]. The consequence of these obstacles is a deficit of training data. Since tagging performance is dependent on the domain of the training data, creating new domain-specific corpora is important for improving tagging accuracy [3].

Recently, research has focused on creating new speech and language data using crowdsourcing. Much like outsourcing, which sends work to other companies or overseas, crowdsourcing sends work to be solved by the public at large [4]. Companies and researchers have found that crowdsourcing over the Internet can be used to solve problems quickly and cheaply. Many papers have addressed using crowdsourcing for creating speech and language data ([5]-[23], discussed later in Chapter 2), but no research has addressed using crowdsourcing to create POS tagged corpora. Since most crowdsourced annotation has involved novice, or untrained annotators, research has focused on simple annotation tasks.

In this work, we determine whether untrained annotators can accurately perform the task of part of speech annotation. Since POS tagging is a difficult task and untrained annotators may not have much linguistic knowledge, we developed a novel Java Applet to assist users during annotation. Although previous research has developed POS tagging interfaces [24], they are not designed for untrained annotators, nor do they incorporate the tagging guidelines. The Applet, referred to as the Interactive Tagging Guide (ITG) follows the Penn Treebank tagset and simplifies the task of annotation. We use the Penn Treebank tagset because of its popularity, limited size, and the availability of written tagging guidelines.

We use the ITG to simplify the task of POS tagging into a series of multiple choice questions. To tag a word, a user is presented a series of questions which guides them to the correct tag. The advantages of this system are a complete elimination of tags from the user interface, the incorporation of the tagging guidelines into the system, and a simplification of the task by breaking apart the job of annotation. In addition, we developed the ITG to automatically tag certain common words and punctuation, thus reducing the annotators' workloads. We developed the ITG for users with a basic knowledge of nouns, verbs, adjectives, and adverbs. By using this guide, we eliminate the need for annotators to be familiar with the Penn Treebank tagset or be trained for POS tagging.

We evaluate the performance of untrained annotators by deploying the ITG on Mechanical Turk, an online crowdsourcing marketplace. We have each word annotated five times to obtain multiple judgments for each word. We compare the aggregated

annotations of Workers on Mechanical Turk to an expert annotation. Furthermore, we evaluate the performance of two expert annotators using the ITG. We then compare the inter-annotator agreement between the untrained annotators using the ITG, the expert annotators using the ITG, and the expert, gold-standard annotation.

1.1 Organization

We present an overview of crowdsourcing for creating speech and language data in Chapter 2. We include an analysis of the incentives used to motivate participation, as well as the recent use of Mechanical Turk to perform crowdsourced annotation. An overview of Mechanical Turk is presented in Chapter 3. We explain the interface, the experience for both Requesters and Workers, and examine the demographic information of Workers on Mechanical Turk. The development and details of the ITG are presented in Chapter 4. Appendices A through C contain the questions used by the ITG, as well as a list of words and punctuation automatically tagged by the ITG.

The experimental setup and results for untrained annotators are presented in Chapters 5 and 6. We then compare the results of untrained annotators to that of experts using the ITG in Chapter 7. We then address the viability of our system and conclusions in Chapter 8.

Chapter 2: Previous Work

The concept of acquiring large amounts of machine learning training data from untrained users can be attributed first to Stork's Open Mind Initiative [5]. Using the Open Source methodology as his motivation, Stork's initiative described a method for training intelligent systems involving three different types of participants. The "domain expert" would be responsible for producing the supervised learning algorithms needed to process large amounts of training data [6]. The "tool developers" would create a user interface, as well as collect and error check the incoming training data. Finally, the "e-citizens", general users of the Internet, would interact with the user interface and provide data which would be fed back to train the system. For example, an optical character recognition system would send images of various characters to e-citizens, who would respond with the character they believed each image represented. The data would then be aggregated, screened for outliers, and then returned as training data to improve classification.

2.1 Importance of Incentives

A critical factor for the success of an Open Mind system, or any crowdsourcing system, is a strong incentive for e-citizens (users) to participate. Stork proposed that e-

citizens could be incentivized to participate by such mechanisms as lotteries, discounts, and frequent flier miles [6]. However, he strongly believed that all users would be interested in the progression of the system. Stork writes,

> "Just as parents delight in watching the cognitive development of their child, so too would contributors be excited to see an Open Mind common-sense reasoning system develop its "understanding" of the world." [6]

In contrast to Stork, von Ahn and Dabbish [7] used entertainment as the incentive for users to participate. They created "the ESP game", a competitive game designed to create labeled images. In the game, two players list properties (such as contents and characteristics) of randomly assigned images. The goal for each user is to guess what properties the other user listed. When they guess correctly, both move on to a new image and are awarded points. While the user is motivated to earn points, they are in fact creating a list of labels for each image.

Other systems, such as Verbosity [8] and Open Mind Word Expert [9] used various degrees of a game structure to incentivize Internet users to participate. Verbosity was another two player game where one player writes clues to help the other player guess the identity of a random object. The output of the game was used to form a database of common sense facts [8]. Open Mind Word Expert was designed to directly collect word sense information. Unlike the ESP game and Verbosity, Open Mind Word Expert did not obscure the nature of the task within a game. Instead, users were given the ability to track their contributions and work their way to the "Hall of Fame" [9].

A key issue with entertainment-based crowdsourcing is that it relies on the task to be entertaining enough to attract enough users [25]. Using a financial incentive to

participate can be used to overcome this obstacle. Launched in 2005, Amazon's Mechanical Turk [26] is an online marketplace where businesses and researchers post small "human intelligence tasks" (referred to as "HITs") for users, or "Workers", to complete. Workers volunteer to complete these HITs, and receive compensation in return. Since HITs are often extremely short, the compensation received for completion is small—often only a few cents. The details of Mechanical Turk are discussed in Chapter 3.

2.2 Studies Involving Mechanical Turk

Several recent studies have used Mechanical Turk as a method of rapidly obtaining cost-effective annotated data. Su [10] used Mechanical Turk for the tasks of entity resolution of hotel records, as well as attribute extraction of age, product brands and models. Each experiment consisted of 300 tasks with each task, such as extracting the brand from a product description, assigned to its own HIT [10]. Each task was completed by three to five unique workers. A "voted answer" was calculated for each task using a simple majority wins voting scheme. If a task obtained a voted answer (2 out of 3 or 3 out of 5), then Workers who submitted that answer were compensated one cent. Workers who submitted a different answer, or worked on a task that did not produce a voted answer were not paid [10]. Data acquisition took between 15 and 18 days and returned accuracies of between 69% and 97%, as compared to a gold standard.

Snow [11] successfully used Mechanical Turk to obtain annotations for five tasks: affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation. Each task (HIT) was completed by ten Workers. The complexity of each task was an important consideration. Snow attempted to minimize the length of task descriptions and provide examples whenever possible. Tasks were chosen that only required a multiple choice selection or a numerical answer within a fixed range [11]. Snow reported especially high rates of annotation, from 90 to 1700 labels per hour, as well as low rates of pay, from 333 to 3500 labels per dollar. In addition, the accuracy for each task was approximately 90% or higher—much higher than those reported by Su. An accuracy of 99.4% was reported for Word Sense Disambiguation, where the only disagreement between the gold standard and Mechanical Turk stemmed from an error in the gold standard. Most notably, Snow demonstrated that for affect recognition, only an average of four Worker labels were necessary to approximate an expert level quality label for a given item [11].

In 2010, the North American Chapter of the Association of Computational Linguistics (NAACL) held a workshop on Creating Speech and Language Data with Amazon's Mechanical Turk [27]. Hoping to expand the work of Snow, the workshop focused on what tasks untrained annotators could accomplish, as well as what new research is possible when the cost of new data is reduced [12]. To answer these questions, 35 researchers were provided $100 to spend on developing their choice of annotated data using Mechanical Turk. Thirty-four papers were accepted on subjects

including machine translation, speech, opinion mining, prepositional phrase attachment, and word sense disambiguation.

Since much of the Mechanical Turk workforce can be assumed to be untrained annotators, many tasks utilized common knowledge or abilities. Finin [13] and Lawson [14] dealt with the task of named entity recognition. Lawson simplified the task by requesting the user to label all entities and state whether they were named or unnamed. Unnamed entities were simply removed after submission. Marge [15] analyzed Worker transcriptions of spontaneous speech and found the aggregated transcription's word error rate to be comparable to that of previous work. Heilman [16] asked Workers to evaluate the quality of automatically generated questions on a simple numeric scale. Yano [17] investigated the ability of Workers to identify political bias in sentences and to identify if it was liberal or conservative.

Several studies utilized the multilingual ability of the international and diverse workforce on Mechanical Turk to create data for machine translation systems [12]. Denowski [18] studied the evaluation of English paraphrases of Arabic sentences. Ambati [19], Gao [20], and Bloodgood [21] explored the task of creating parallel corpora in a variety of languages. A common issue in these studies was the use of answers obtained from online automated machine translation systems. Checks to prevent this form of cheating were necessary to stop this behavior.

No previous research has attempted to create a similar POS tagged corpora using Mechanical Turk. The most related study was by Jha [22], where Mechanical Turk was used to construct a prepositional phrase attachment corpus from informal or blog text.

Building upon the work of [23], Jha developed a heuristic-based system to extract prepositional phrases and their possible attachment points. The ability to list the possible attachment points allowed them to simplify the task by reducing the space of possible options. As a result, they transformed the complex task of identifying prepositional phrases and their attachments into a series of basic multiple choice questions. Five workers answered each question and the majority (or plurality) answer was selected. The overall accuracy was reported as 84.48% [22]. Jha noted that accuracy increased with greater Worker agreement. For example, questions with unanimous agreement had an accuracy of 97.43% [22].

2.3 Contribution

In this work, we build upon the previous work of creating language data with Mechanical Turk and attempt the novel task of creating a POS tagged corpus using crowdsourcing. One of the main questions of the NAACL 2010 workshop was whether complex annotation can be done by untrained annotators [12]. This work provides further evidence to the range and complexity of tasks which can be accomplished using untrained annotators on Mechanical Turk.

In this work, we follow the methodology of previous work, such as Snow [11] and Jha [22], as well as incorporate some new concepts. We follow the approach of task simplification by having Workers interface with the Interactive Tagging Guide. As in [22], we use this guide to convert a complicated task to a series of multiple choice

questions. Unlike the semi-automated system of [22], which is responsible for question construction, our system is also responsible for automatically labeling the parts of speech of approximately one-third of all tokens. This method reduces the complexity, workload, and cost of annotation. We also follow the convention of using multiple Workers to answer each question. In this study, five Workers are used to label each word. Finally, we follow the common convention of comparing the labels of Workers to a gold standard to evaluate accuracy.

Chapter 3: Amazon's Mechanical Turk

Mechanical Turk [26] was launched in 2005 by Amazon as an online service where users earn small payments for the completion of short tasks. The idea behind Mechanical Turk is that, while computers are powerful at certain tasks, there are still many tasks at which they perform poorly. Many of these tasks can be completed by humans, often with little to no effort [28]. To signify the use of humans in these tasks, Mechanical Turk adopted the motto "artificial artificial intelligence". The service's name is derived from an 18th century chess-playing "automaton" which secretly was controlled by a person inside a hidden compartment [12].

Mechanical Turk is an on-demand marketplace where often menial tasks can be completed at a fraction of the cost of paying an employee to do so. On Mechanical Turk, tasks are referred to as "human intelligence tasks", or HITs. HITs are posted by "Requesters", who are often businesses or researchers. Once a Requester has posted HITs to Mechanical Turk, they can be selected and completed by "Workers"—users who sign up to earn money completing HITs.

The number of HITs available varies at any time. At the time of writing, there were approximately 184,000 HITs available on Mechanical Turk [26]. Compensation for completing a HIT can be as little as one cent to as much as several dollars, although most do not pay very much. Of the 2175 "HIT groups" (A HIT group is a collection of similar

HITs) available on Mechanical Turk at the time of writing [26], 1360 pay less than fifty cents per HIT.

3.1 Requesters on Mechanical Turk

Requesters are provided a great amount of flexibility in creating HITs on Mechanical Turk. HITs can be created using an online graphical user interface, command line tools, or an API that can automatically incorporate results into an application. HITs can contain a variety of multimedia and interactive elements, such as images, sounds, videos, and Java Applets. HITs may also be hosted on an external website for unlimited control over the task design. A Requester may decide to limit the types of Workers than can complete their HITs by assigning qualifications. Common qualifications include geographic location, HIT approval rate, and number of HITs completed. Requesters can also require Workers to satisfactorily complete a qualification test before working on a HIT.

The results of a HIT are sent to Mechanical Turk, not the Requester, for evaluation. Once the results have been sent to Mechanical Turk, they may be reviewed by the Requester. Workers are only identified by a unique alphanumeric Worker ID to protect their anonymity. Requesters have the ability to approve or reject the results of any HIT based on the answers received. If a HIT is approved, a payment is sent via Amazon Payments to the Worker. If a HIT is rejected, then the Worker is not paid—even though the Requester still has access to their answer. A Requester may choose to

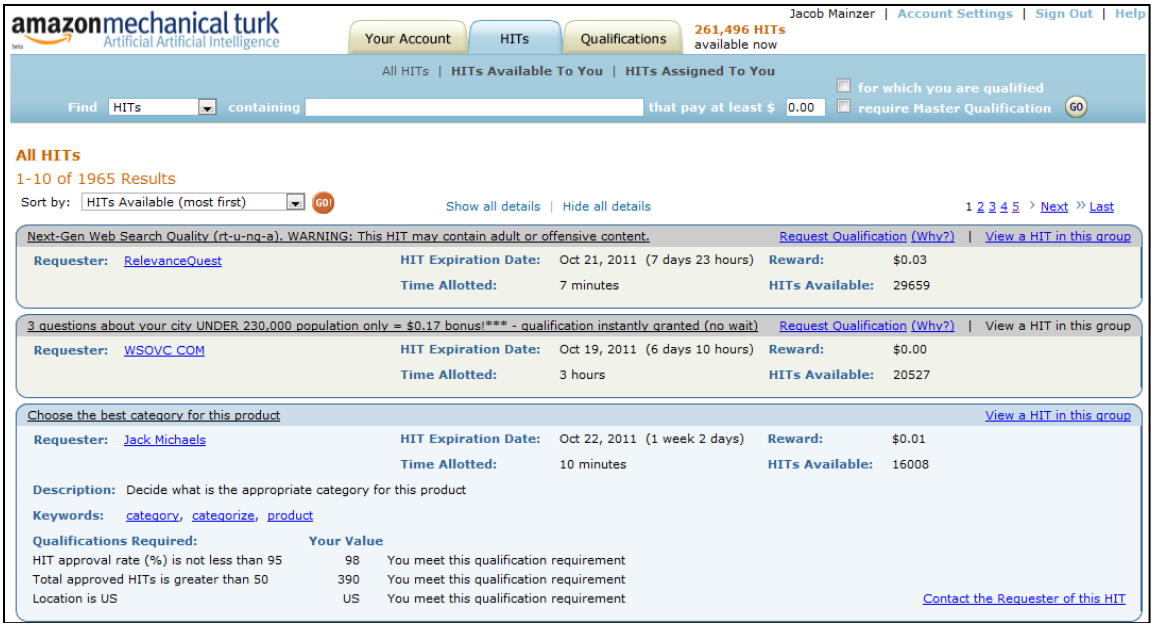

Figure 1: Three HIT groups posted on Mechanical Turk. This page is where Workers can search for and select HITs that they wish to complete. Note that some HITs have qualifications which a Worker must meet in order to participate.

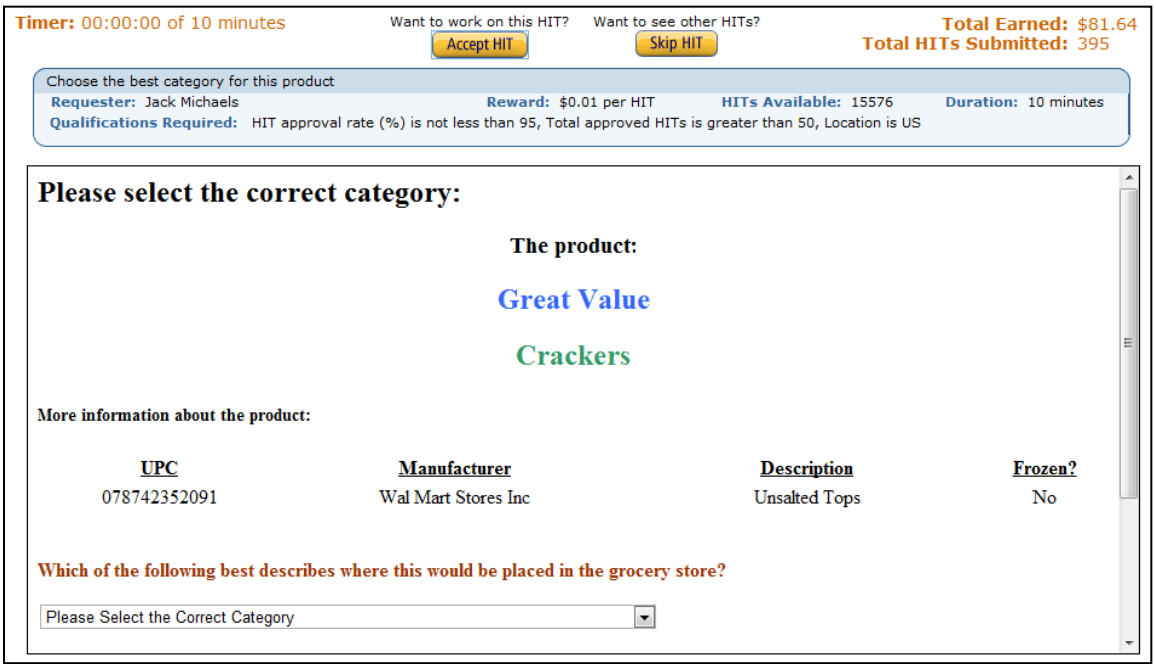

Figure 2: The HIT interface prior to accepting the HIT. A preview of the task is shown and the Worker can choose to accept the HIT.

acknowledge exceptional answers by paying a bonus amount of their choosing. Amazon charges a 10% fee on all payments and bonuses, with a minimum fee of $0.005 per HIT.

3.2 Workers on Mechanical Turk

Workers can search and choose any posted HIT that they wish to complete from a searchable list, as shown in Figure 1. When a Worker chooses a HIT, they "accept" it and the HIT appears in a window on their browser. The interface prior to accepting a HIT is shown in Figure 2. If a Worker choses to stop working on a HIT, they can "return" the HIT at no penalty. Once a Worker has completed a HIT, they must "submit" their work to Mechanical Turk, where it is reviewed by the Requester.

There are over 400,000 workers registered with Mechanical Turk [29]. The workforce is highly global, with 44% of all Workers located outside of the United States. Mechanical Turk has an especially large presence in India, where 36% of the workforce is located [29]. Considering the often low rate of pay, one might suspect that most Workers are young, uneducated, low income earners looking to "make ends meet". Surprisingly this is not the case, especially with U.S. Workers. American Workers are reported, on average, to be 35 years old and 55% of those surveyed have a Bachelor's degree or higher [29]. Forty-one percent reported their annual household income as $50,000 or higher. Only 14% of U.S. Workers reported that they used the income from Mechanical Turk to "sometimes or always" make ends meet [29]. American Workers are also almost twice as likely to be female than male.

The prosperity of many Workers has not stopped the comparisons of Mechanical Turk to a "digital sweatshop" [30]. The average hourly salary on Mechanical Turk is only $1.97 [29], well below the U.S. minimum wage. In addition, Requesters can reject HITs without reason and without the ability for recourse. Rejected HITs not only deny payment to the Worker, but also negatively affects their "HIT approval rate"—a qualification used on many HITs. Since Mechanical Turk has no method of rating Requesters, unaffiliated sites such as Turker Nation [31] and Turkopticon [32] have been developed to discuss and review Requesters. These websites allow Workers to check the reputation of Requesters prior to accepting a HIT. As a result, a positive reputation on these websites can be useful for a Requester to attract attention to their HITs.

Chapter 4: System Development

4.1 Motivation and Considerations

We designed a system to assist untrained annotators on Mechanical Turk to accurately and quickly POS tag words using the Penn Treebank tagset. The Penn Treebank tagset (Table 1) was used because of its limited size, as well as its popularity. The tagset contains 36 POS tags and 12 tags for symbols and punctuation [2]. The size is a sharp reduction from previous tagsets. The Brown Corpus tagset, from which many other tagsets have been derived [33], utilizes 87 tags. However, due to the ability to form compound tags, the Brown Corpus actually contains 179 different tags [33]. In the effort to recognize more grammatical distinctions, The Lancaster/Oslo-Bergen Corpus of British English (LOB) created a larger set of 132 non-compound tags. Other tagsets are even larger, such as the London-Lund tagset which contains almost 200 tags [33]. The number of tags in these tagsets would increase the challenge of untrained annotation—perhaps prohibitively. Consequently, the Penn Treebank tagset was chosen for this study.

An important consideration in the design of this task was whether the system would involve "corrective tagging", users correcting preexisting tags, or "manual tagging", users tagging words from scratch. Experiments using expert annotators during the creation of the Penn Treebank demonstrated that manual tagging took twice as long as

1	CC	Coordinating Conjunction	25	TO	To
2	CD	Cardinal number	26	UH	Interjection
3	DT	Determiner	27	VB	Verb, base form
4	EX	Existential 'there'	28	VBD	Verb, past tense
5	FW	Foreign word	29	VBG	Verb, gerund/present participle
6	IN	Preposition/subordinating conjunction	30	VBN	Verb, past particle
7	JJ	Adjective	31	VBP	Verb, non-3rd person singular present
8	JJR	Adjective, comparative	32	VBZ	Verb, 3rd person singular present
9	JJS	Adjective, superlative	33	WDT	wh-determiner
10	LS	List item marker	34	WP	wh-pronoun
11	MD	Modal	35	WP$	Possessive wh-pronoun
12	NN	Noun, singular or mass	36	WRB	wh-adverb
13	NNS	Noun, plural	37	#	Pound sign
14	NNP	Proper noun, singular	38	$	Dollar sign
15	NNPS	Proper noun, plural	39	.	Sentence-final punctuation
16	PDT	Predeterminer	40	,	Comma
17	POS	Possessive ending	41	:	Colon, semi-colon
18	PRP	Personal pronoun	42	(	Left bracket character
19	PRP$	Possessive pronoun	43	)	Right bracket character
20	RB	Adverb	44	"	Straight double quote
21	RBR	Adverb, comparative	45	‘	Left open single quote
22	RBS	Adverb, superlative	46	“	Left open double quote
23	RP	Particle	47	’	Right close single quote
24	SYM	Symbol	48	”	Right close double quote

Table 1: The Penn Treebank Tagset. Note that the possessive pronoun tag was incorrectly listed as PP$ in [2].

correction and resulted in 75% more inter-annotator disagreement [2]. The mean error rate for correction was also 25% lower than the mean error rate for manual tagging [2].

Although correction appears to be a faster, more accurate form of annotation, this method is problematic when applied to untrained annotators. Untrained annotators are

not familiar enough with the tagset to make meaningful corrections. Furthermore, unscrupulous or hurried annotators may not attempt to make corrections and consequently report incorrect tags as being correct. As a result, we chose to have users perform manual annotation.

Pure manual annotation is also problematic for untrained annotators. Generating the correct tag from even a relatively small tagset is difficult—even for trained annotators. Annotators on the Penn Treebank reported a learning curve of a just under a month, working 15 hours a week [2]. Furthermore, the annotators had graduate training in linguistics. Untrained annotators cannot have a learning curve and must be able to instantly annotate accurately. Asking an untrained annotator to do the same task as an expert is not reasonable and would not result in expert-level annotations.

4.2 The Interactive Tagging Guide: Amend, Assist, and Automate

Since untrained annotators would have difficulty performing POS tagging, we alter the task by following three main key points: amend, assist, and automate. We amend the task of POS tagging by breaking apart the process of choosing a tag into a series of multiple choice questions. In cases of difficult words and tags, we assist the user with special word-specific questions to help select the correct tag. Finally, we automate the task by automatically tagging 24 different parts of speech. Following these three key points, we developed a Java Applet, referred to as the Interactive Tagging Guide (ITG) to specifically facilitate POS tagging by untrained annotators. The ITG was

CC (11)	DT (6)	MD (2)	PDT	POS	PRP (3)
PRP$ (2)	TO	WDT (3)	WP (2)	WP$	WRB (2)
#	$	.	,	:	(
)	"	‘	“	’	”

Table 2: Twenty-two tags which are automatically tagged by the ITG. The number after certain tags signifies the number of individual words which cannot be automatically tagged. Those words have their own word-specific questions.

created from the written guidelines of the Penn Treebank tagset. The guidelines for the Penn Treebank tagset are described in detail in [34]. The 32-page document includes a description of each tag, information on how to tag difficult cases, and a list of problematic words and collocations.

We designed the ITG to be the interface used by Workers on Mechanical Turk. Upon loading, The ITG receives a string of words (typically a sentence) and a sequence of binary digits. Each binary digit represents a word from the input string and tells the system whether a word should be annotated or not. This allows control over which words a user should annotate and is useful if a Worker is being asked to annotate only a fraction of a sentence. In a web environment, these input parameters can be passed to an Applet by using JavaScript.

Before the user begins annotation, the ITG checks which words can be automatically tagged, or “auto-tagged”. We found that the annotation of 24 tags could be automated or semi-automated (automated with the exception of specific individual words). Table 2 lists these tags. Punctuation and symbol tags make up 12 of the 24 tags. Instances of each punctuation and symbol type can be easily identified and assigned a

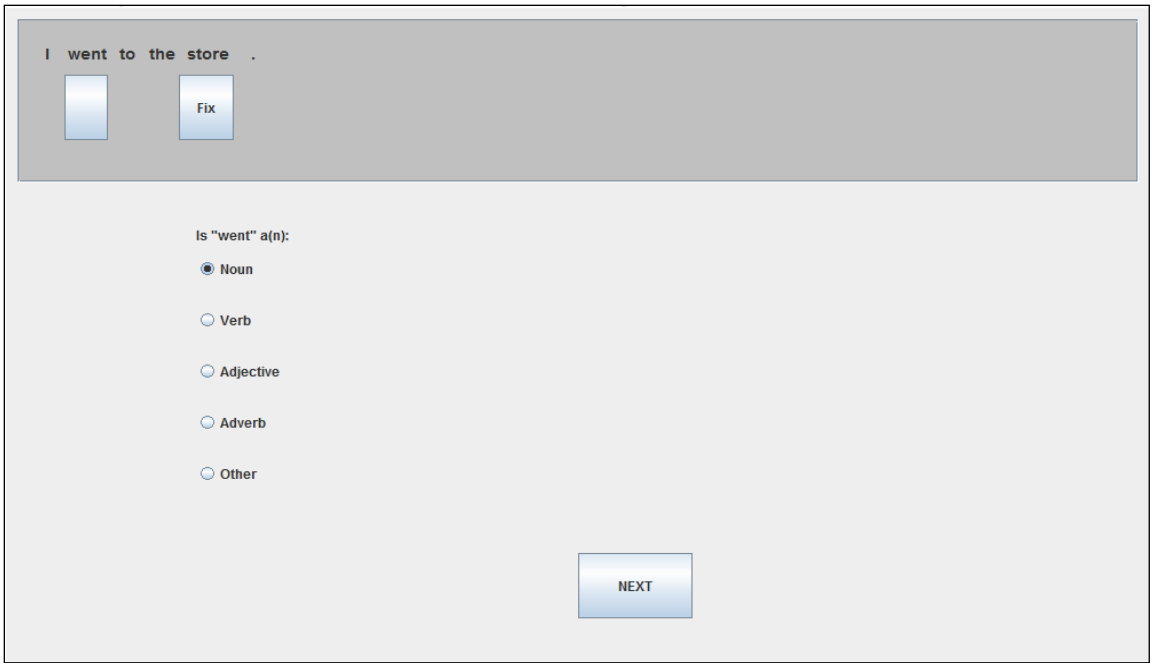

Figure 3: The interface used for annotation. The opening question is shown for the word 'went'. The sentence is shown in the dark grey box at the top of the image. A button below a word selects the word for annotation. Words with no button underneath ('I', 'to', 'the', and '.') have been auto-tagged.

tag. The remaining 12 parts of speech are small, closed sets. Many of the words in these sets are given in [34]. The ITG includes a list of these words and their parts of speech.

Special cases of collocations and context are also automatically handled by the ITG. The Penn Treebank tagging guidelines specifies 37 collocations that result in specific annotations. For example, 'due' is always tagged as an adjective, JJ, except in the collocation "due to", when it is tagged as a preposition, IN [34]. The ITG identifies these collocations and annotates them automatically. Furthermore, the ITG uses context to auto-tag certain tags. For example, a determiner that precedes another determiner or possessive pronoun should be labeled as a predeterminer, PDT [34]. The ITG

automatically annotates predeterminers in this context. Auto-tagged words and collocations are included in Appendix C.

Once all possible words have been autotagged, the user may begin their annotation. An example of the interface for annotation is shown in Figure 3. We amended the task of manual POS annotation into a series of guided, multiple choice questions. The user may choose to annotate any word in any order. For each word, the user is asked a series of questions in order to select the correct tag. These questions form a decision tree, with each decision node represented by a question, and each leaf node represented by a POS tag.

Our goal was to create questions that required a minimum of linguistic knowledge. We assumed that our annotators had a basic understanding of what a noun, verb, adjective, and adverb are, as well as how to identify them in their standard uses. A qualification test could be used to verify whether a Worker possesses a basic understanding of these parts of speech. In this experiment, however, we did not use qualification tests. Sixteen tags are members of one of these four basic parts of speech. Furthermore, we identified five additional tags which we believed that untrained annotators could easily label. These include foreign words, interjections, list item markers, cardinal numbers, and symbols. Cardinal numbers are autotagged by the ITG when they consist of only digits (e.g. 123.24). However, cardinal numbers written as words (e.g. ‘six’) must be labeled by the annotator. An example for labeling a singular, common noun (NN) is shown in Figures 4 and 5. The complete question set for the ITG is included in Appendix A.

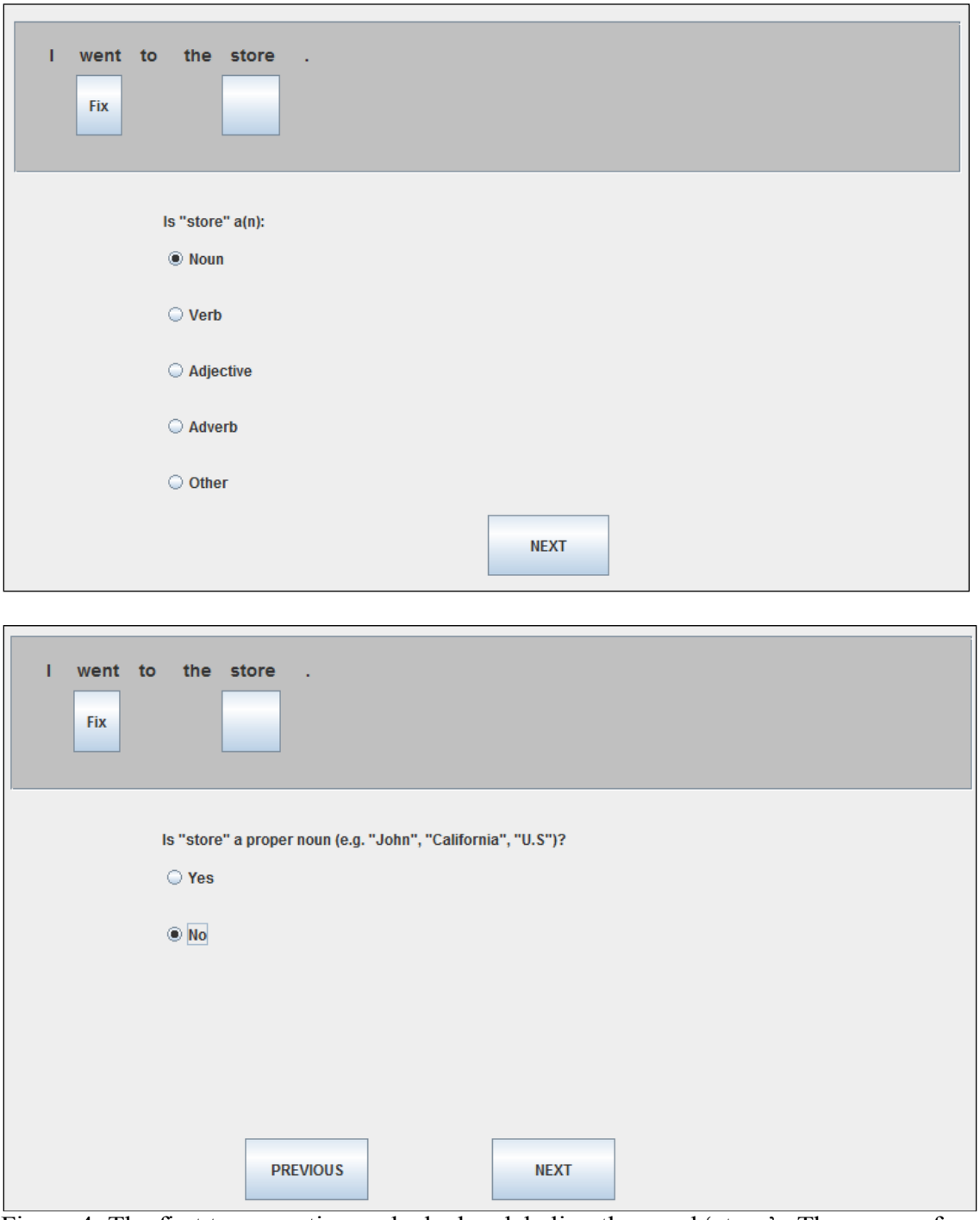

Figure 4: The first two questions asked when labeling the word ‘store’. The answer for each question has been selected.

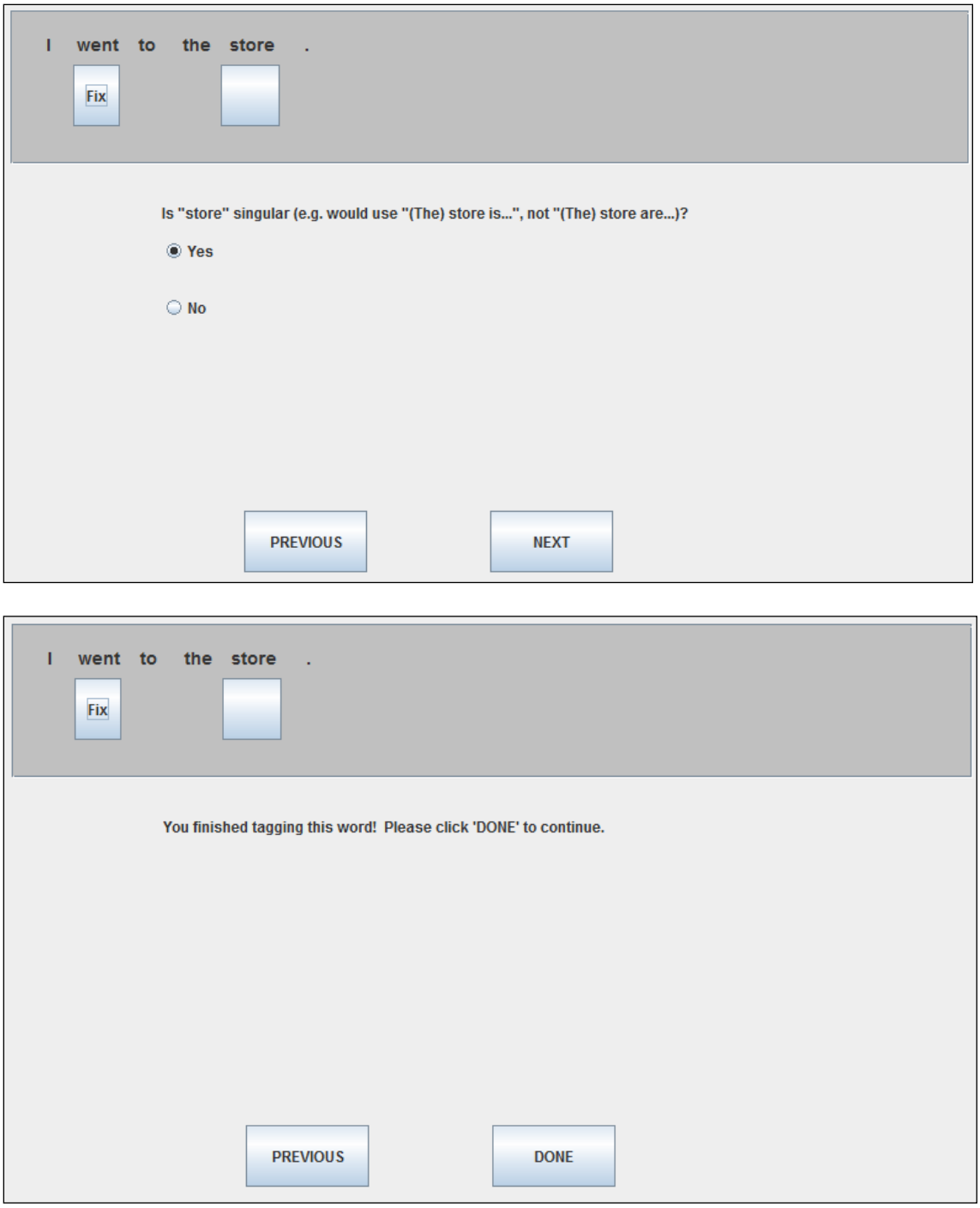

Figure 5: The last question and success prompt for labeling the word ‘store’. Note that a POS tag is not shown at any point in the labeling process or after a tag has been selected (the final tag was NN). As a result, an untrained annotator can label a word without any knowledge of the tagset.

Occasionally, certain parts of speech may be easily confused. This may especially be the case when the Penn Treebank tagging guidelines do not follow what an untrained annotator may generally expect. Examples of confusable parts of speech are including in the [34] as “problematic cases”. For example, the tagging guidelines state that nouns that are used as modifiers should be labeled as nouns, not adjectives. As a result, in the phrase “wool sweater”, ‘wool’ should be labeled as a noun. Annotators may easily mislabel the word as an adjective. The problematic cases section of [34] includes tests and examples to assist annotators in selecting difficult tags.

In order to provide additional assistance to untrained annotators, the ITG includes tests to decide between difficult parts of speech. Many of the tests were taken from [34]. The ITG test questions are included in Appendix A. In the example of the “wool sweater”, if the annotator initially selects that ‘wool’ is an adjective (using the question seen in Figure 3), then they will be asked the following question:

Could "wool" be a Noun or Verb?

- **It could be a noun**
- **It could be a verb**
- **No, it's definitely an adjective**

If the annotator believes that ‘wool’ could be a noun, then they are presented the following test:

Is "wool":

- **able to be modified by a degree adverb like "really" or "very"? (e.g. "A really fun trip.")**
- **a proper noun that serves the role of an adjective? (e.g. "I bought Chinese food.")**
- **An adjective or noun that cannot be modified by a degree adverb (e.g. "This is a dark red", "red" cannot be modified--*"This is a dark very red")**

If the annotator selects the third answer, then ITG labels 'wool' as a common noun. As a result of the test questions included in the ITG, an annotator can be guided from the wrong tag to the correct tag.

The ITG also assists annotators with labeling specific words by using 51 word-specific questions. The Penn Treebank tagging guidelines include a list of words which may be problematic to tag [34]. For example, the guidelines state that the word 'about' should be labeled as an adverb if it used to mean 'approximately'; otherwise it should be labeled as a preposition. The guidelines for these words were converted to questions that apply only to the specific words. For example, when tagging the word 'about', the user does not see the usual starting question (Figure 3), but rather the following question:

Is "about" being used to mean "approximately"?

- **Yes**
- **No**

If the user selects 'Yes', as in the case of "The man was about 6 feet tall", then the word is tagged as an adverb. Otherwise, the word is tagged as a preposition. A word-specific question is also used to differentiate the existential use of 'there' ("There is a

problem.") from the adverbial use of there ("The problem is there."). The existential use of 'there' has its own tag, EX. All 51 word-specific questions are given in Appendix B.

A combination of auto-tagging, question answering, and word-specific questions are used to annotate 46 out of 48 parts of speech (including punctuation). The two remaining parts of speech, prepositions and particles, require additional attention. Prepositions are a closed set. However, a putative preposition may be labeled as a particle or adverb. As a result, simple auto-tagging is not sufficient.

The ITG identifies prepositions automatically and asks the user a series of preposition-specific questions. Prepositions are identified by searching a manually constructed set of prepositions. If a word is identified as a putative preposition, then the annotator is asked a series of questions to determine if the word should be labeled as a preposition, particle, or adverb. These questions are derived from tests in the "problematic cases" section of [34].

The ITG also has several other POS-specific questions. If an annotator labels a word with the suffix "-er" or "-est" as an adjective, then they are asked if the word has a comparative or superlative meaning, respectively. If an annotator labels a word with the suffix "-ing" as a verb, then they are indirectly asked if the word is a gerund. If the annotator labels that same word as an adjective, they are automatically presented a test to determine if the word should be tagged as an adjective or gerund. All upper-case words have their own question to determine whether they are part of a name or title.

In the development of the Interactive Tagging Guide, we removed all references to the Penn Treebank tagset. Even after a word has been tagged, the user does not see the

tag they have chosen. By completely eliminating all tags from the ITG, we transformed the task of POS tagging so that an annotator needs no knowledge of the tagset or its guidelines. Instead, the tagset guidelines are built into the system.

A summarization of the features of the Interactive Tagging Guide is presented below:

- Follows the Penn Treebank tagset guidelines as given in [34].
- Automatic tagging (auto-tagging) of 24 parts of speech and 37 collocations.
- A multiple-choice question/answer system used to annotate 21 parts of speech.
- Fifty-one word specific questions used to assist labeling problematic words.
- Test questions used to differentiate easily confusable parts of speech.
- Automatic detection of prepositions and preposition-specific questions used to differentiate prepositions, particles, and adverbs.
- Complete elimination of all tags from the user interface.
- Assumes only general knowledge of nouns, verbs, adjectives, and adverbs.

Chapter 5: Experiment

We ran an experiment on Mechanical Turk to test the hypothesis that untrained annotators can accurately label parts of speech. In this experiment, we had Workers on Mechanical Turk label 50 English sentences from Wikipedia. Workers used the Interactive Tagging Guide to perform all annotation. The untrained annotators' results were compared to those of expert annotators.

5.1 Dataset

We created a corpus of 50 English sentences randomly taken from Wikipedia[1]. Each sentence was taken from a random Wikipedia article. We selected only Wikipedia articles that were written in English (although we accepted the use of foreign words) and at least five sentences long. Each sentence was randomly selected from among the first five sentences of a Wikipedia article. Two sentences contained slight grammatical errors and were left unchanged.

The corpus was initially tagged by the Stanford POS tagger [35] [36], a publically available tagger that uses the Penn Treebank tagset. The tags were then hand-

[1] While we would have preferred to use the Wall Street Journal corpus for this experiment, we were unable to due to copyright issues.

Tag	Count	Percentage
CC	38	3.3%
CD	33	2.8%
DT	119	10.2%
EX	1	0.1%
FW	0	0.0%
IN	141	12.1%
JJ	67	5.7%
JJR	3	0.3%
JJS	2	0.2%
LS	0	0.0%
MD	2	0.2%
NN	161	13.8%
NNS	44	3.8%
NNP	194	16.6%
NNPS	9	0.8%
PDT	0	0.0%
POS	4	0.3%
PRP	20	1.7%
PRP$	9	0.8%
RB	28	2.4%
RBR	1	0.1%
RBS	0	0.0%

Tag	Count	Percentage
RP	0	0.0%
SYM	0	0.0%
TO	22	1.9%
UH	0	0.0%
VB	19	1.6%
VBD	31	2.7%
VBG	14	1.2%
VBN	25	2.1%
VBP	8	0.7%
VBZ	34	2.9%
WDT	3	0.3%
WP	1	0.1%
WP$	0	0.0%
WRB	3	0.3%
.	50	4.3%
,	47	4.0%
:	5	0.4%
(	12	1.0%
)	12	1.0%
“	2	0.2%
”	2	0.2%
Total	1166	100.0%

Table 3: The distribution of parts of speech in the experimental corpus, as defined by the expert annotation. Table includes the number of a given tag in the corpus, as well as the percentage of the entire corpus. Punctuation that was not used in the dataset is not included in this table.

corrected by three expert annotators: the author and two linguistics graduate students. The mean inter-annotator agreement was extremely high, at 98.0%. We found that the original tags created by the Stanford tagger were extremely accurate. Only 4.5% of the tags were changed during hand correction.

The corpus is composed of 1036 words and 130 punctuation marks. The distribution of the 1166 tags in the corpus is shown in Table 3. The corpus contains a

Words to Manually Label/HIT	Cost/HIT (with Fees)	Number of HITs	Total Cost (5 Workers/HIT)
1 Word	$0.015	11	$0.825
2 Words	$0.025	9	$1.125
3 Words	$0.035	12	$2.10
4 Words	$0.045	7	$1.575
5 Words	$0.055	140	$38.50
	TOTAL	**179**	**$44.125**

Table 4: Distribution and cost of HITs by the number of words Workers are required to manually annotate.

large number of nouns (35.0%), especially proper nouns (17.4%). The corpus has no instances of seven different POS tags and five punctuation tags. Most notably, while there are 141 instances of prepositions (IN), there are no instances of particles (RP). The expertly annotated corpus can be found in Appendix D.

5.2 Experimental Setup

We divided the task of annotation into a series of HITs. Each HIT asked a Worker to label a fraction of a sentence. We created 179 HITs, with each HIT containing no more than five words which required manual annotation. Of the 1166 words and punctuation in the corpus, 373 (32.0%) are auto-tagged by the ITG. The remaining 793 words must be manually annotated.

Each HIT was completed by five unique Workers in order to obtain multiple labels for each word. Workers could complete multiple HITs and were paid at the rate of one cent per manually annotated word. No compensation was given for auto-tagged words. Unlike [10], we approved all Worker results, regardless of whether they

submitted the majority answer. For each HIT, a half-cent fee was paid to Mechanical Turk. The total cost of compensation and fees for all HITs was $44.125. A breakdown of the distribution and cost of HITs for this experiment is shown in Table 4.

We had two qualifications for Workers to participate in this experiment. In order to target high quality, thoughtful Workers, we required Workers to have a 95% HIT acceptance rate. In addition, we required all Workers to be located in the United States. We limited the HITs to US Workers in an effort to target native English speakers. It is unknown what fraction of the Worker population satisfies these requirements.

Upon starting a HIT, Workers first were shown an informed consent page, including a task description, compensation amount, risks, benefits and contact information. After agreeing to participate in the experiment, Workers were directed to the task's webpage. The webpage was hosted by Amazon S3, external to Mechanical Turk. The parameters of the HIT, including Worker ID, Assignment ID, sentence string, and words to annotate were passed to the webpage as URL parameters. On the task webpage were short instructions, a brief reminder of what nouns, verbs, adjective, and adverbs are, and the ITG Applet. At startup, the Applet received the sentence and the desired words to annotate as parameters from the website. If a Worker did not have Java installed, they were provided a link to download the necessary software. The task webpage is show in Figure 6.

After reading the instructions, Workers used the ITG to label the words requested by the HIT. Although we estimated that each HIT would take about 5 minutes to finish,

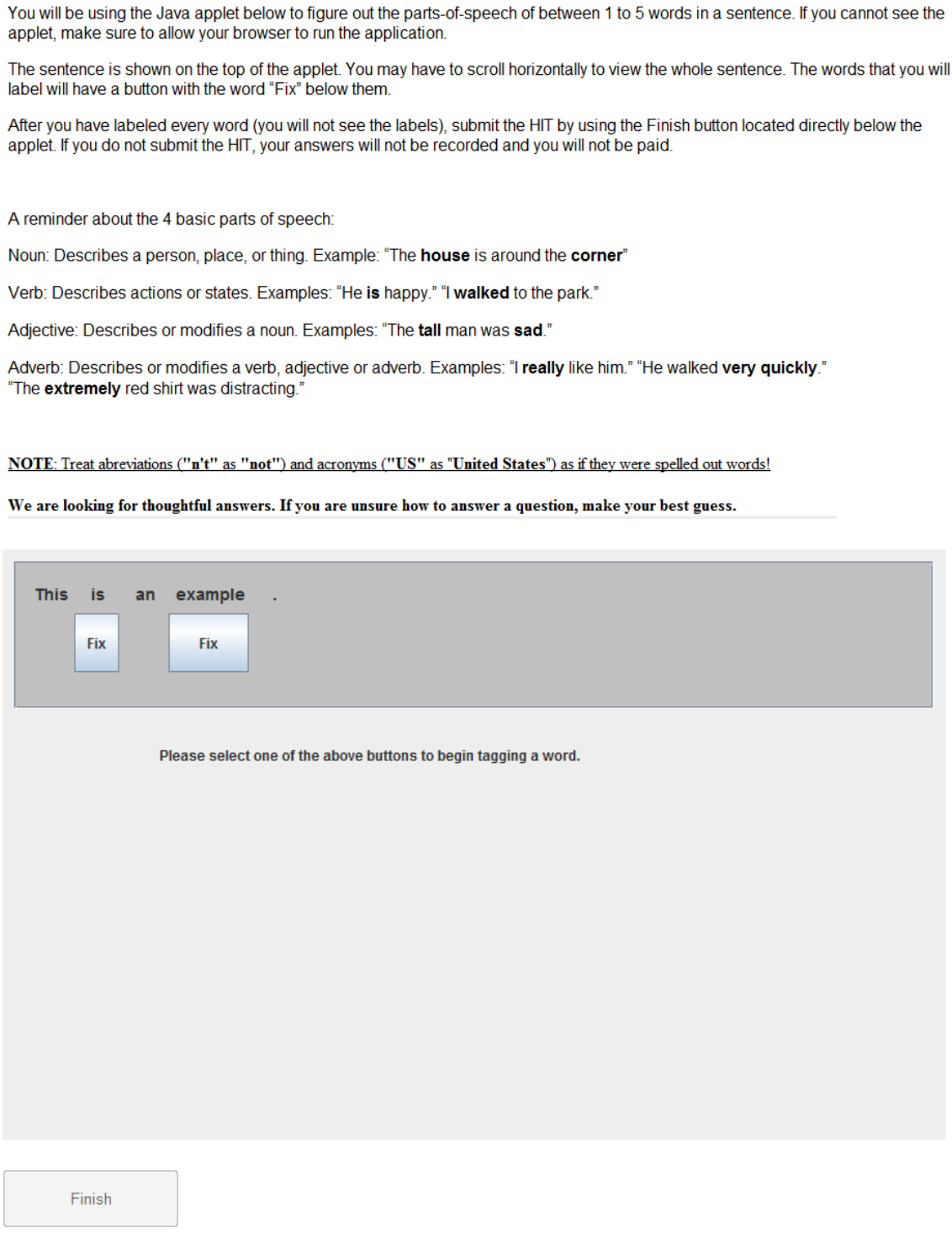

Figure 6: The task webpage used for completing HITs. The ITG is embedded in the bottom half of the webpage. Note that Workers saw this page within the window shown in Figure 2.

Workers were given one hour to complete and submit the HIT. If a Worker did not submit their results within one hour, the HIT was returned and any work completed was thrown out. After the user finished annotating with the ITG, the Applet used JavaScript to transfer the results from the applet to an invisible submission form on the task webpage. When Workers pressed the "Finish" button on the webpage, their answers were sent to the Mechanical Turk servers for evaluation.

Chapter 6: Experimental Results

In this experiment, we had 50 sentences from Wikipedia POS tagged by Workers on Mechanical Turk. Each of the 1166 words and punctuation were tagged by five unique Workers. The accuracy of these individual annotations is 85.9% for both manually and auto-tagged labels and 79.3% for manually tagged labels only. To improve accuracy, a final tag is determined by a simple plurality voting scheme. The tag with the most number of votes is selected as the final answer. There are two answer distributions where no final tag can be determined: the 2-2-1 distribution, where there is agreement on two different tags, and the 1-1-1-1-1 distribution, where each worker selects a different tag. In these instances, no tag is assigned.

The experimental results are summarized in Table 5. Using a plurality voting system, Workers on Mechanical Turk are able to POS tag with 84.6% accuracy, as compared to the expertly generated tags. The 373 auto-tagged labels have an accuracy of 100%. As a result, the tags produced by the ITG system, as a combination of manual and automatic annotation, have an accuracy of almost 90%. The mean inter-annotator

	All Tags	Manually Labeled Tags
Amount Correct	1044	671
Total	1166	793
Percentage Correct	**89.5%**	**84.6%**
95% Confidence Interval	**+/- 1.8%**	**+/- 2.5%**

Table 5: Experimental results for final tags.

Agreement	Number of Tags	Percent of All Tags	Accuracy
5-0 (auto-tagged & manual)	792	67.9%	98.4%
5-0 manually labeled	419	35.9%	96.9%
4-1	185	15.9%	87.0%
3-2	101	8.7%	68.3%
3-1-1	43	3.7%	72.1%
2-2-1	37	3.2%	*0.0%
2-1-1-1	6	0.5%	66.7%
1-1-1-1-1	2	0.2%	*0.0%

Table 6: The distribution of tags by Worker agreement. The accuracy of each distribution is shown on the right. *The 2-2-1 and 1-1-1-1-1 distributions do not produce a final tag and therefore have an accuracy of 0%.

agreement rate is 82.7% for all tags, and 74.6% for manually tagged labels only. The agreement rate for Workers is far lower than the 98.0% agreement rate of our expert annotators. However, recall that corrective tagging, as was done for the expert annotation, results in higher inter-annotator agreement [2]. In addition, previous work has shown that noise should be expected in data acquired using Mechanical Turk [12].

While inter-annotator disagreement is high, over half (419/793) of the manually annotated tags have unanimous Worker agreement. When including the auto-tagged labels, which always have unanimous agreement, 67.9% of the tags have unanimous agreement. The distribution of tags by Worker agreement is shown in Table 6. As in [22], accuracy improved with greater Worker agreement. This suggests that voter agreement can serve as a confidence measure for tags. Tags with unanimous Worker agreement have an accuracy of 98.4%. Furthermore, the decrease in accuracy seen in agreements below 4-1 suggests the use of a hybrid strategy, where only 5-0 and 4-1 tags are used, and all tags below a 4-1 distribution are sent for expert annotation.

Question	Times asked	Correct %
start	2160	88.4%
noun-start	946	95.8%
prop-noun	986	97.3%
common-noun	845	98.2%
verb-start	543	77.7%
verb-ing	65	81.5%
VB-VBP-test	80	47.5%
VBN-VBD-test	233	82.4%
VBG-NN-test*	8	37.5%
VBG-JJ-test*	7	57.1%
adj-comp	15	73.3%

Question	Times asked	Correct %
adj-sup	9	100.0%
JJ-NN-test*	70	70.0%
JJ-VBN-test*	75	74.7%
adv-start*	66	83.3%
two-prep-test	15	53.3%
standed-IN-test	31	35.5%
IN-RP-test	549	73.2%
RP-RB-test	0	N/A
All word specific questions	235	76.6%
CD-start*	3	100.0%
upper-case	1060	91.1%

Table 7: The Worker accuracy for each question asked by the ITG. The number of times asked does not include situations where the Worker has already made an uncorrectable mistake. An asterisk denotes that the question was a test question which could correct a previous incorrect decision. The RP-RB-test was only used after a Worker made an uncorrectable error. See Appendix A for a complete listing of all questions used by the ITG.

The 2-2-1 and 1-1-1-1-1 distributions account for 3.4% of all tags. Since neither of these distributions has a plurality answer, tags with this distribution are not assigned a final tag. Eliminating tags with these distributions from the results improves the accuracy to 92.6% overall, and 89.0% for manually labeled tags only. In addition, 32 out of 37 tags with a 2-2-1 answer distribution had the correct tag selected by one set of two workers. Consequently, randomly selecting a tag with two votes would yield approximately 16 correct tags out of 37. Using a random selection process in case of a tie improves the overall accuracy to 90.9%.

The Worker performance on each question is listed in Table 7. A correct answer is defined as an answer which accurately represents the word being annotated. For example, labeling "store" as a noun (Figure 5) would be a correct response to the initial **START** question. The response is correct even if future questions result in the Worker assigning an incorrect tag to the word. An incorrect answer is an answer which incorrectly represents the word being annotated. Using the previous example, labeling "store" as an adjective would be an incorrect response to the **START** question because "store" is not an adjective. The answer is incorrect even though the Worker could still label the word correctly (as a common noun, NN) using the **JJ-NN-TEST**. As a result, a Worker may answer questions incorrectly and still return a correct tag. Questions asked after a Worker makes an uncorrectable error are not included in the performance statistics, since neither answer may be correct. Questions that were never used or which an accuracy measurement cannot be defined are not included in Table 7. See Appendix A for a complete listing of all questions used by the ITG.

The assumption that Workers had basic knowledge of nouns, verbs, adjectives, and adverbs appears to be well-founded. Workers were able to correctly identify these four basics parts of speech 88.4% of the time. Workers were highly accurate in answering questions about nouns, with each question correctly answered over 95% of the time. Workers had trouble answering questions about verbs. Workers were only correct 47.5% of the time when answering whether a verb was VB (base form verb) or VBP (non-3rd person singular present verb). The question, which was taken directly from the tagging guidelines of [34], actually performed worse than chance. Workers also had

Tag	Count	Recall
CC	38	94.7%
CD	33	84.8%
DT	119	100.0%
EX	1	100.0%
FW	0	N/A
IN	141	80.1%
JJ	67	89.6%
JJR	3	100.0%
JJS	2	100.0%
LS	0	N/A
MD	2	100.0%
NN	161	88.8%
NNS	44	93.2%
NNP	194	99.0%
NNPS	9	77.8%
PDT	0	N/A
POS	4	100.0%
PRP	20	100.0%
PRP$	9	100.0%
RB	28	50.0%
RBR	1	100.0%
RBS	0	N/A

Tag	Count	Recall
RP	0	N/A
SYM	0	N/A
TO	22	100.0%
UH	0	N/A
VB	19	26.3%
VBD	31	100.0%
VBG	14	92.9%
VBN	25	44.0%
VBP	8	75.0%
VBZ	34	73.5%
WDT	3	100.0%
WP	1	100.0%
WP$	0	N/A
WRB	3	66.7%
.	50	100.0%
,	47	100.0%
:	5	100.0%
(	12	100.0%
)	12	100.0%
“	2	100.0%
”	2	100.0%

Table 8: The recall for each POS, as calculated for the final tags selected by a plurality voting scheme.

difficulty answering questions related to prepositions and particles.

Workers issues with labeling verbs and prepositions are also apparent from the final tag recall rates, as shown in Table 8. Workers were able to correctly tag the six verb parts of speech only 69.5% of the time. Workers had the most issues identifying verbs in their base form (VB), with only 5 out of 19 instances correctly labeled. Worker error on verb parts of speech accounted for 33% of the total Worker error. Although

prepositions had a recall of 80.1%, they accounted for 23% of the total Worker error. The large error stems from Workers erroneously labeling prepositions as particles. Although the expert annotation contains no particles, the Workers' annotation contains 18 particles. Of the words which did not obtain a plurality answer (2-2-1 or 1-1-1-1-1 distribution), 20.5% of those were prepositions.

While Workers had some difficulty labeling verbs and prepositions, Workers were extremely accurate labeling nouns. Workers correctly labeled 93.9% of all nouns. Furthermore, Workers correctly labeled 99.0% of all singular proper nouns (NNP). Workers only made two errors labeling singular proper nouns—one of which was an expert error. These results suggest that Workers were able to easily recognize and label nouns.

Workers also correctly labeled 90.3% of all adjectives and correctly identified all superlative and comparative adjectives. Workers were only able to identify only 51.7% of adverbs. Workers labeled 9 out of 29 adverbs as adjectives, accounting for 64% of Worker error for adverbs. While adverbs were mistaken for adjectives, no adjectives were incorrectly labeled as adverbs. This suggests that Workers tended to label modifiers as adjectives.

6.1 Bias

The conversion of the written Penn Treebank tagging guidelines to the Interactive Tagging Guide may result in bias in ITG annotations. Although the ITG follows the

Word	Gold Std Tag	Worker Tag
But	CC	IN
that*	IN	WDT
Firing	NN	VBG
Rest	VB	VBP
long*	RB	JJ
Encourage	VB	VBP
even*	RB	JJ
West	NN	JJ
Australian	JJ	NNP
Of	IN	RP
yet*	RB	CC
Panamanian	JJ	NNP
One	CD	PRP

Table 9: Words with unanimous Worker agreement but did not match the gold standard tag. Four words, designated with asterisks, were correctly labeled by Workers but have an incorrect gold standard tag.

tagging guidelines, annotation may be affected by the multiple-choice questions used by the ITG. Bias resulting from the ITG is difficult to distinguish from Worker error. We present two methods of estimating the bias caused by the ITG.

We assume that words which all five Workers improperly annotated to be the result of bias. The core assumption is that if no bias is in effect, then at least one Worker should be able to properly annotate the word. If no Workers properly label the word, then the ITG may be leading them to the wrong tag. In our experiment, 2.9% of manually annotated tags had no Workers provide the correct answer. Using this method, we estimate the bias of the ITG to be 2.9%.

Additionally, we observe 13 instances where all five Workers chose the same tag, but did not choose the same tag as the expert annotation. These words are shown in

Table 9 and represent 1.6% of all manually labeled words. Of these 13 words, four words were determined to be correctly labeled by the Workers and incorrectly labeled in the gold standard. The remaining nine words represent 1.1% of all manually labeled words and represent a baseline measure of bias.

Our final measure of bias is derived from the effect of test questions on Worker annotations. Test questions are used to differentiate difficult tags and correct Workers when they select the wrong tag. Test questions are defined as having the ability to change the POS tag contrary to the previous assertions by the Worker. Recall the example of the "wool sweater", where the Worker initially asserts 'wool' is an adjective, but a test question corrects the Worker and labels the word as a noun. There are six questions which meet the definition of a test question.

While a test question can guide a Worker from the wrong answer to the right answer, it can also incorrectly guide the Worker away from the right answer. We derive a bias measure from the number of final tags where test questions caused the annotation to be incorrect. In these situations, Workers were guided away from the correct tag by the test questions. If the test questions had not changed one or more Workers' answers, then the tag would have been correct. In our experiment, 11 final tags are incorrect due to the use of test questions. Using this measure, we estimate the ITG bias to be 1.4% for manually annotated tags. Combined with the previous bias estimate, we approximate the bias of the ITG to be 1.4-2.9% of manually annotated tags.

A valuable test question should help more Workers than it misleads. Since test questions are optional, a question with an overall negative effect should be removed from

Question	Caused User Error	Corrected User Error	Did Not Cause User Error	Did Not Correct User Error	Net effect	# Incorrect Final Tags due to Question-Caused Error	# Correct Final Tags due to Correction of User Error
VBG-NN-test	5	2	1	0	-0.4	0	0
VBG-JJ-test	3	2	2	0	-0.2	1	0
CD-start	0	0	3	0	N/A	0	0
adv-start	11	5	50	0	-0.4	2	1
JJ-NN-test	10	41	8	11	0.60	1	11
JJ-VBN-test	16	4	52	3	-0.6	7	1
Total	45	54	116	14	0.1	11	13

Table 10: Effect of test questions on Worker tags and Final tags. The net effect of each question is calculated as ([corrected user error] – [caused user error])/([corrected user error] + [caused user error]). A question with a negative net effect caused more errors than it fixed. A question with a positive net effect fixed more errors than it caused.

the ITG. Table 10 shows the positive and negative effects of the six test questions. Five out of six test questions had an overall negative effect on Worker performance. For example, the **JJ-VBN-test** corrected 4 individual Worker tags, but also changed 16 tags from the right POS to the wrong POS. Omitting this test would result in a net gain of 12 correct individual tags and 6 final tags (using the plurality voting scheme).

We measured the net effect of each test question as the average effect on the number of correct individual tags due to a question changing the POS of a Worker's answer. For example, each time the **JJ-VBN-test** is used to change a Workers' tag from JJ to VBN or VBN to JJ, the number of correct individual tags is reduced, on average, by 0.6. A negative net effect value means that the test question lowers tag accuracy, while

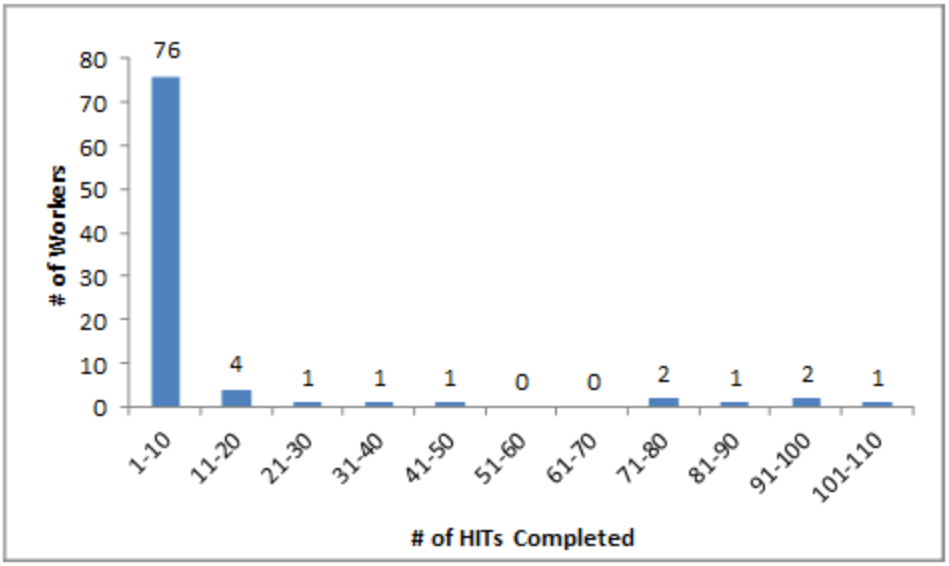

Figure 7: Distribution of Workers by the number of HITs completed.

a positive net effect value means the question improves tag accuracy. The net effect of each test question is included in Table 10.

Despite the negative impact of 5 out of 6 test questions, the net effect of all test questions combined was positive. The test questions are responsible for improving the number of correct tags by two tags. This is due to the **JJ-NN-test**, which has a strong positive net effect. The **JJ-NN-test** results in a net gain of 10 correct individual Worker tags. Removing all negative test questions while keeping the **JJ-NN-test** improves the final tag accuracy to 85.6% for manually annotated tags and 90.2% for all tags.

6.2 Worker Statistics

A total of 89 Workers participated in this experiment. Figure 7 shows the

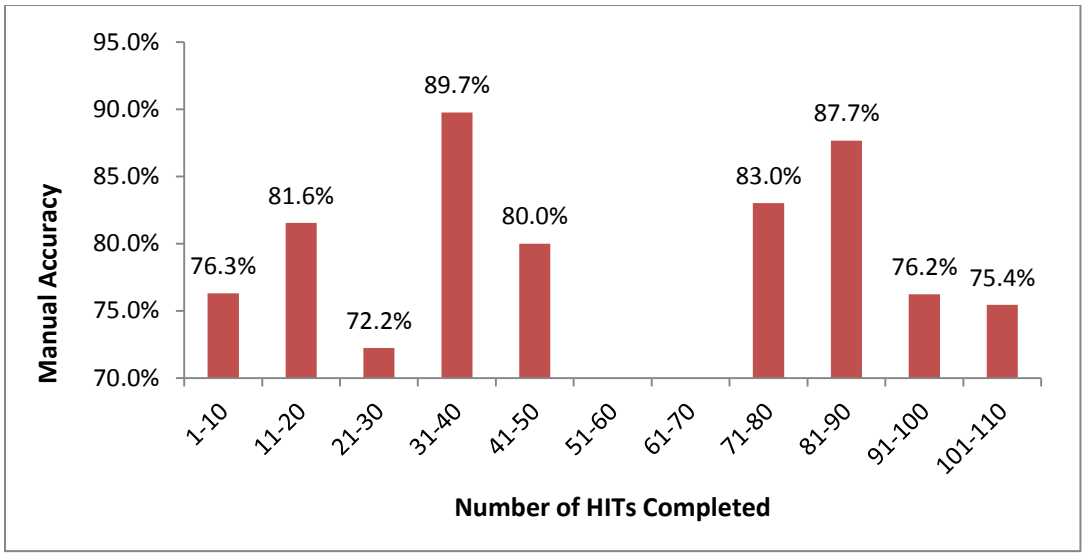

Figure 8: The accuracy of Workers by the number of HITs completed.

distribution of Workers by the number of HITs they completed. 84.3% of all Workers completed less than 10 HITs. As in [11] a small minority of Workers completed most of the HITs. In our experiment, 60% of all HITs were completed by just six Workers. The disparity in HITs completed is the result of Workers being allowed to complete as little or as many HITs as they desire.

Since the majority of results are obtained from a small number of Workers, the final accuracy is heavily dependent on the accuracy of these Workers. As a result, it would be advantageous for these Workers to have the highest tagging accuracy. Figure 8 shows the accuracy of Workers by the number of HITs they completed. The accuracies of Workers are not dependent on the number of HITs they complete. The most prolific Worker, who completed 110 HITs, had a lower accuracy than the average Worker who completed between 1-10 HITs.

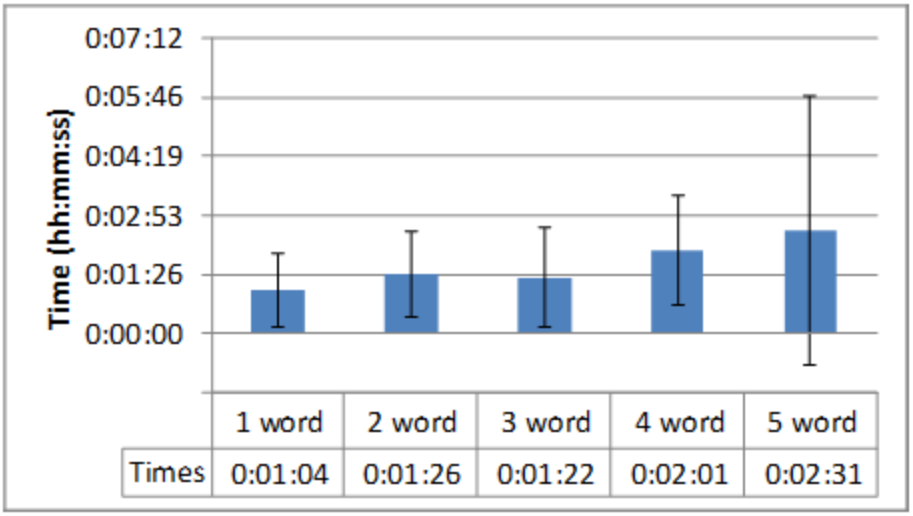

Figure 9: Average Worker time to complete an individual HIT, for HITs of 1-5 words. Error bars show standard deviation of completion times.

Workers were able to complete the HITs at an approximate rate of 30 seconds per manually labeled word. Figure 9 shows the average time to complete each of the five different HITs. The times include the time it took Workers to read the informed consent document. There was a large amount of variation in the amount of time Workers spent completing HITs. For the task of manually annotating 5 words, Workers took between 37 minutes and 15 seconds to complete the task. Overall, Workers completed all the HITs for this experiment within 135 hours (5.6 days) of posting. This results in an overall annotation speed of 43.2 individual tags per hour.

Chapter 7: Supplemental Experiment and Analysis

In the previous experiment, we demonstrated that untrained annotators can POS tag with an accuracy of 89.5%. To determine whether this accuracy is on par with an expert, we have two expert annotators use the ITG to annotate the same sentences as the Workers in the previous experiment. Each expert, referred to as Expert A and Expert B, annotated 50 sentences using the ITG in the Mechanical Turk interface. If the ITG allows untrained annotators to POS tag with the accuracy of an expert, then Experts A and B should have the same tagging accuracy as the Workers.

Experts A and B, using the ITG, were able to POS tag with 93.1% and 89.8%, respectively (Table 11). Only Expert A was able to POS tag with a significantly higher accuracy than the Workers. A result is significant ($p = 0.05$) if the 83% confidence intervals do not overlap [37] [38]. Expert A was also significantly more accurate than Expert B. Since the difference in accuracies between Expert B and the Workers is not significant, we conclude that the Workers were able to perform expert level annotation.

	Accuracy
Expert A	93.1%
Expert B	89.8%
Workers	89.5%
POS Tagger	95.5%

Table 11: Accuracy of the four annotations, as compared to the gold standard annotation (corrected output of the POS tagger).

	Expert A	Expert B	Gold Std.	Workers	POS Tagger
Expert A	-	91.8%[1]	93.8%[1,3]	91.3%[1,2]	90.2%[1]
Expert B	91.8%[1,2]	-	90.9%[2]	90.3%[1]	88.1%[1]
Gold Std.	93.8%[2]	90.9%[1]	-	92.6%[2]	95.4%
Workers	91.3%[1]	90.3%[1,2]	92.6%[2,3]	-	89.0%[1]
POS Tagger	90.2%[1]	88.1%[2]	95.4%[1]	89.0%[1]	-

Table 12: Inter-annotator agreement rates between all five annotations. Rates are for the 1127 words which all annotations include. Words that the Workers did not assign a label for (2-2-1 or 1-1-1-1-1 distributions) are not included. Differences between agreements in the same column with the same superscript are not statistically significant (p=0.05).

Note that the difference in accuracy between Expert A and the POS tagger can be accounted for by the bias estimate given in Chapter 6.

7.1 Analysis

We have referred to the expert correction of the POS tagger as the "gold standard" annotation. However, the gold standard does not necessarily follow the tagging guidelines. We discovered seven errors in the gold standard, all of which were correctly labeled by the Workers. Further undetected errors in the gold standard may exist due to the bias of corrective tagging. Since the gold standard annotation starts with the output of the POS tagger, we believe that the gold standard may be biased towards the output of the tagger.

To investigate the relationship between the annotations of the Workers, Experts A and B, the gold standard, and the POS tagger, we compare the inter-annotator agreement rates between all of these annotations in Table 12. Since 3.4% of the final Worker annotation is not assigned a tag, the inter-annotator agreement between Workers and

other annotations will be artificially low. Consequently, we analyze the inter-annotator agreement only for tags for which all annotations provide an answer.

Table 12 shows Workers have the same level of agreement with the gold standard (92.6%) as Experts A (93.8%) and B (90.9%). This further strengthens the case that Workers using the ITG are able to annotate at an expert level. It is interesting to note that users of the ITG do not agree with each other more than they agree with the gold standard. This suggests that, despite using the ITG, the Experts and Workers all have disagreement on different areas of the corpus. Although the ITG follows the Penn Treebank guidelines, it still requires user input, and that leads to disagreement and error.

The inter-annotator agreement between the POS tagger and the gold standard (95.4%) is significantly higher than the agreement between the POS tagger and all other annotations. Annotations using the ITG do not have the same level of agreement with the POS tagger. This suggests our assertion that the gold standard is biased towards the POS tagger is correct. Since the POS tagger does not necessarily follow the tagging guidelines, bias towards the tagger may result in annotations which do not conform to the tagging guidelines. Consequently, calculating the accuracy of the Worker and Expert A and B annotations from the gold standard may improperly decrease the reported accuracy of these annotations. Furthermore, calculating the accuracy of the POS tagger based on the gold standard may over-represent the accuracy of the POS tagger.

Chapter 8: Viability and Conclusions

It is important to consider the viability of creating POS tagged corpora on Mechanical Turk. We have already shown that Workers on Mechanical Turk can POS tag with the accuracy of an expert, but is Mechanical Turk cost and time effective? Is using Mechanical Turk cheaper than hiring an expert annotator? We analyze the case of creating a million tag corpus below.

In the previous experiment, Workers created 1166 tags for $44.125, or 3.8 cents per tag. The cost for a million tag corpus would be approximately $38,000—a substantial amount. In comparison, 3 experts who can correct 3000 tags per hour [2] and are paid $10 per hour can create a million tag corpus for $11,800 (including 60 hours of training). In addition, assuming the rate of annotation from our experiment, creating a million word corpus on Mechanical Turk would take 13 years to complete. Three experts working 15 hours a week could create the same corpus in approximately 22 weeks. Under these assumptions, these results call into question the viability of Mechanical Turk for large scale POS annotation. However, we believe that the rate of annotation is not constant. Furthermore, the cost and time estimates of expert annotation do not include the cost of finding, recruiting and retaining annotators. These factors increase the difficulty in starting the process of annotation and are avoided by using the ITG on Mechanical Turk.

The cost of annotation on Mechanical Turk can easily be reduced by 30% without cutting pay or affecting accuracy by eliminating redundant data. In order to select a final tag, generally three Workers have to agree on the same tag (2-1-1-1 distribution being the exception). Therefore, once three Workers agree on a tag, no other Workers need to label the word. For example, in the case of unanimous agreement, obtaining the last two tags is not necessary. In our experiment, 65.6% of all manually labeled tags could be labeled by only three Workers. In other words, these tags had complete annotator agreement for the first three Workers. Only 15.4% of all manually labeled tags needed to be labeled by five Workers. Halting annotation of a word once three Workers agree reduces the cost of annotation by 30%. Furthermore, it also should reduce the time of annotation by 30%, since there are 30% less HITs for Workers to complete. The cost of a million tag corpus is reduced to $26,600—still over twice as expensive as using experts. Further work should investigate the effect of lowering the cost per word below the rate of one cent per word used in this experiment.

We believe that the rate of annotation on Mechanical Turk would increase for the creation of a larger corpus. A larger annotation task would produce more HITs, which may attract more Workers to the project. Furthermore, a larger project would attract more attention from Requester review sites, such as Turker Nation [31] and Turkopticon [32]. Positive feedback could direct even more Workers to the task. Further word of mouth and the opportunity to complete large amounts of HITs should further drive participation. In order to expand the workforce, future research may also look at

loosening the qualification requirements, such as being located in the U.S., for Workers to participate.

Future work on the question set and auto-tagging may also improve performance and cost. We demonstrated that 5 out of 6 test questions had a negative impact on performance. Modifying or removing these questions should improve performance. In this experiment, we did not attempt to expand the auto-tagging of words beyond those included in the tagging guidelines. Future work may look at attempting to expand auto-tagging to all words with only one tag. Auto-tagging all words that do not need user input would reduce the cost and effort in creating new corpora. Finally, future work should investigate incorporating a tie-breaking or weighted voting scheme to resolve 2-2-1 and 1-1-1-1-1 distributions.

8.1 Conclusions

We find that untrained annotators on Mechanical Turk can POS tag with high accuracy. The use of a plurality voting scheme and the Interactive Tagging Guide allows Workers to label at near 90% accuracy. Furthermore, we find that Experts using the ITG annotate at the same approximate level of accuracy as the aggregated results of the Workers. Although the Stanford POS tagger has a higher agreement with the gold standard than the Workers, we observe a bias in the gold standard towards the POS tagger output. We believe the bias of the gold standard overstates the accuracy of the gold standard and understates the accuracy of annotations using the ITG.

We believe this work expands the range of tasks which untrained annotators may be used. The key features of the ITG: assist, amend, and automate, serve as a template for further extending untrained annotation in POS tagging and other speech and language tasks. Perhaps new tagsets could forgo creating paper tagging guidelines such as [34] and construct interactive guides with built-in guidelines. Furthermore, future versions of the ITG could utilize the international makeup of Mechanical Turk to perform POS tagging in new languages and domains. It is clear that the limits of crowdsourcing for speech and language data creation have not been reached. Future work should attempt to further utilize crowdsourcing as a means of obtaining complex and valuable annotations.

References

[1] E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics*, vol. 21, no. 4, 1995, pp. 543-565.

[2] M.P. Marcus, B. Santorini, and M. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, 1993, pp. 313-330.

[3] M. Dredze, J. Blitzer, P. Talukdar, K. Ganchev, J. Graca, and F. Pereira, "Frustratingly Hard Domain Adaptation for Dependency Parsing," *Proceedings of EMNLP-CoNLL 2007 Shared Task*, 2007.

[4] J. Howe. (2006, June). *The Rise of Crowdsourcing* [Online]. Available: http://www.wired.com/wired/archive/14.06/crowds.html

[5] D. G. Stork, "The Open Mind Initiative," *IEEE Intelligent Systems & Their Applications*, vol. 14, no.3, 1999, pp. 19-20.

[6] D.G. Stork, "Character and Document Recognition in the Open Mind Initiative," *Proc. Int'l Conf. Document Analysis and Recognition* (ICDAR '99)*,* IEEE Computer Society Press, Los Alamitos, Calif., 1999.

[7] L. von Ahn and L. Dabbish, "Labeling Images with a Computer Game," In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '04), 2004, pp. 319-326.

[8] L. von Ahn, M. Kedia, and M. Blum, "Verbosity: A Game for Collecting Common-Sense Facts," In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '06)*,* 2006, pp. 75-78.

[9] T. Chklovski and R. Mihalcea, "Building a Sense Tagged Corpus with Open Mind Word Expert," *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*, 2002, pp. 116-122.

[10] Q. Su, D. Pavlov, J. Chow, W.C. Baker, "Internet-Scale Collection of Human-Reviewed Data," *Proceedings of the 16th international conference on World Wide Web* (WWW '07), 2007, pp. 231-240.

[11] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng, "Cheap and Fast---But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '08), 2008, pp. 254-263.

[12] C. Callison-Burch and M. Dredze, “Creating speech and language data with Amazon's Mechanical Turk,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[13] T. Finin, W. Murnane, A. Karandikar, N. Keller and J. Martineau, “Annotating Named Entities in Twitter Data with Crowdsourcing,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[14] N. Lawson, K. Eustice, M. Perkowitz, and M. Yildiz, ”Annotating large email datasets for named entity recognition with Mechanical Turk,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[15] M. Marge, S. Banerjee, and A. Rudnicky, “Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[16] M. Heilman and N.A. Smith, “Rating Computer-Generated Questions with Mechanical Turk,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[17] T. Yano, P. Resnik, N.A. Smith, “Shedding (a Thousand Points of) Light on Biased Language,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[18] M. Denkowski, H. Al-Haj, and A. Lavie, “Turker-Assisted Paraphrasing for English-Arabic Machine Translation,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[19] V. Ambati and S. Vogel, “Can Crowds Build Parallel Corpora for Machine Translation Systems?” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[20] Q. Gao and S. Vogel, “Semi-Supervised Word Alignment with Mechanical Turk,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[21] M. Bloodgood and C. Callison-Burch, “Using Mechanical Turk to Build Machine Translation Evaluation Sets” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[22] M. Jha, J. Andreas, K. Thadani, S. Rosenthal, and K. McKeown, “Corpus Creation for New Genres: A Crowdsourced Approach to PP Attachment,” In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[23] S. Rosenthal, W.J. Lipovsky, K. McKeown, K. Thadani, and J. Andreas, “Towards Semi-Automated Annotation for Prepositional Phrase Attachment,” In *Proceedings of LREC,* 2010.

[24] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N.A. Smith, "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments," In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, companion volume, 2011.

[25] A. Saini.(2008 May 14). *Solving the Web's Image Problem* [Online]. Available: http://news.bbc.co.uk/2/hi/technology/7395751.stm

[26] Amazon Mechanical Turk (2011 October 12, 6:56pm EDT). Available: https://www.mturk.com/mturk/welcome

[27] NAACL 2010. (2010). *Creating Speech and Language Data with Amazon's Mechanical Turk* [Online]. Available: https://sites.google.com/site/amtworkshop2010/

[28] J. Pontin. (2007 March 25). *Artificial Intelligence, With Help from the Humans* [Online]. Available: http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html

[29] J. Ross , L. Irani , M. Six Silberman , A. Zaldivar , and B. Tomlinson, "Who are the Crowdworkers? Shifting Demographics in Mechanical Turk", *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems* (CHI '10), 2010.

[30] K. Mieszkowski. (2006 July 24). *I Make $1.45 a Week and I Love it*. Available: http://www.salon.com/2006/07/24/turks_3/

[31] Turker Nation. (2011). Available: http://www.turkernation.com/

[32] Turkopticon. (2011). Available: http://www.turkopticon.differenceengines.com/

[33] G. Sampson, "Alternative grammatical coding systems," *The Computational Analysis of English*, R. Garside, G. Leech, and G. Sampson eds., Longman, 1987, pp. 165-170.

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

[35] K. Toutanova and C.D. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger," In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 2000, pp. 63-70.

[36] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

[37] P.C. Austin and J.E. Hux, "A Brief Note on Overlapping Confidence Intervals," *Journal of Vascular Surgery,* vol. 36, 2002, pp. 194-195.

[38] M.E. Payton, M.H. Greenstone, N. Schenker, “Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance?” *J Insect Sci.*, vol. 3, 2003, pp. 34 – 40.

Appendix A: Complete Question Set Used by Interactive Tagging Guide

The following appendix includes the questions used by the ITG. Each question begins with a name (e.g. “start”), followed by the question. The phrase #WORD# represents the current word the user has selected. Following the question are the multiple choice options. The first half of the answer is the choice the user sees. The second half is the either the name of the next question the user will be presented, or the tag the word will be assigned.

start
Is "#WORD#" a(n):
Noun noun-start
Verb verb-proc
Adjective adj-proc
Adverb adv-start
Other other-start

noun-start
Is "#WORD#" a proper noun? (e.g. "John", "California", "U.S")
Yes prop-noun
No common-noun

prop-noun
Is "#WORD#" singular? (e.g. would use "#WORD# is...", not "#WORD# are...)
Yes NNP
No NNPS

common-noun
Is "#WORD#" singular? (e.g. would use "(The) #WORD# is...", not "(The) #WORD# are...)
Yes NN
No NNS

verb-ing
The verb "#WORD#" ends in "-ing". Is this "-ing" a suffix (e.g. "walk-ing", "travel-ing") and not only a verb like "bring" or "sing".
Yes VBG-NN-JJ-test
No verb-start

VBG-NN-JJ-test
Could this verb actually be a Noun or Adjective?
It could be a Noun VBG-NN-test
It could be an Adjective VBG-JJ-test
No, it's definitely a Verb VBG

VBG-NN-test
In the context of this sentence, can "#WORD#"...
be pluralized (e.g. "readings"), modified by an adjective (such as "good", "first"), or is preceded by one or more nouns NN
be modified by an adverb (such as "well") and cannot be pluralized VBG

VBG-JJ-test
Can "#WORD#" either be preceded by a degree adverb, such as "very" or "extremely", or take the prefix "un-" and have the opposite meaning?
Yes JJ
No VBG

other-start
Is "#WORD#" a(n):
Number CD-start
Interjection (a word unrelated to the sentence, e.g. ‘oh’, ‘wow’, ‘yes’, ‘please’) UH
Foreign Word FW
Symbol (NOT a word or abbreviation of English) SYM
Word used to mark items in a list (e.g. “I need <B>a)</B> a pencil and <B>b)</B> a piece of paper") LS

CD-start
Is "#WORD#":
A number-number combination being used like an adjective (a 21-7 score) JJ
A fraction that could be replaced by "double" or "twice" ("one-half the amount") RB
A fraction that modifies a noun and acts like an adjective ("one-half cup") JJ
None of the above CD

adv-start
Is "#WORD#":
A noun that is used like an adverb? (e.g. "He came by Sunday.") noun-start

An adverb with a comparative meaning? (often ends in "-er", e.g. "He runs faster than you.") RBR
An adverb with a superlative meaning? (often ends in "-est", e.g. "He runs the fastest.") RBS
None of the above RB

adj-start
Could "#WORD#" be a Noun or Verb?
It could be a noun JJ-NN-test
It could be a verb JJ-VBN-test
No, it's definitely an adjectiveJJ

adj-comp
Does "#WORD#" have a comparative meaning? (e.g. "Jill is the faster runner")
Yes JJR
NO JJ

adj-sup
Does "#WORD#" have a superlative meaning? (e.g. "Jill is the fastest runner")
Yes JJS
No JJ

JJ-NN-test
Is "#WORD#":
able to be modified by a degree adverb like "really" or "very"? (e.g. "A really fun trip.") JJ
a proper noun that serves the role of an adjective? (e.g. "I bought Chinese food.") prop-noun
An adjective or noun that cannot be modified by a degree adverb (e.g. "This is a dark red", "red" cannot be modified--*"This is a dark very red") common-noun

JJ-VBN-test
Is: "#WORD#":
able to be modified by a degree adverb like "really" or "very" (e.g. "I am very surprised"), or a reference to a state as opposed to an event (e.g. "At that time, I was married") JJ
a reference to an event or action (e.g. "I was married on a Sunday") VBN

verb-start
Is "#WORD#" in:
the basic form? (e.g. “be” for “to be”, “walk” for “to walk”) VB-VBP-test
the present tense, 3rd person singular form? (e.g. "She walks", "He says", "It is") VBZ
some form of the past tense? VBN-VBD-test

the present tense, non-3rd person singluar form (e.g. "I walk", "You are", "They sing", "We look") VBP

VB-VBP-test
If you change the subject of the verb to “he” or “she”, does the verb take an “-s” ending?
Yes VBP
No VB

VBN-VBD-test
If you replaced "#WORD#" with a form of "to see", "to give", or "to know", would that form be: (ignore the change in meaning)
Saw, gave, or knew VBD
Seen, given, or known JJ-VBN-test

prep-test
Does the object of the preposition “#WORD#” immediately follow the word “#WORD#”?
Yes, there is an object right after the preposition (e.g. "I walked under the green table."--the object "the green table" immediately follows the preposition "under") IN-RP-test
No, the object is not right after the prepostion, or there is no object (e.g. "Why don't you come by?"--"by" has no object) stranded-IN-test
The word is not a prepostion start

IN-RP-test
If you replace that object in the sentence with a pronoun (“it”, “them”, "him", "her", etc.), is the sentence still grammatical?
Yes, the sentence is still a proper sentence (e.g. “The dog under the table and chairs barked” becomes “The dog under it barked”) IN
No, the sentence doesn’t make sense. Switching the pronoun and preposition may make the sentence correct. (e.g. “She took off the sticker” becomes *“She took off it”, which can be made correct as “She took it off”) RP

dare-need-test
If you change the subject of the verb to “he” or “she”, does the verb take an “-s” ending?
Yes VBP
No MD

standed-IN-test
Is the object of the preposition "#WORD#" located earlier in the sentence? (e.g. "The table he leaned against was large." The preposition "against" refers to the object "the table" earlier in the sentence)
Yes IN
No RP-RB-test

RP-RB-test
Can you place a manner adverb (an adverb describing how something is done, such as "calmly", "quietly", "easily", "quickly", etc) between "#WORD#" and its associated verb? (e.g. "I passed (quickly) through")
Yes RB
No RP

two-prep-test
Is "#WORD#" more closely associated with the verb before it than the phrase after it?
Yes (e.g. "He hangs out in the coffee shop", the preposition "out" is more associated with the verb "hangs" than the phrase "in the coffee shop") IN
No (e.g. "This has been around for decades", the preposition "around" is more associated with the phrase "for decades" than the verb "to be") RP-RB-test

upper-case
"#WORD#" is capitalized. Is "#WORD#" part of a name? (e.g. "The United States of America", "Wall Street", "Sarah")
Yes, it is part of a name prop-noun
No, it is not part of a name start

Appendix B: Word-Specific Questions of the ITG

The following appendix contains all the word–specific questions used by the ITG. The formatting is the same as in Appendix A.

about
Is "about" being used to mean "approximately"?
Yes RB
No IN

all
Is "all" used like an adverb (e.g. "He traveled all around the city.")?
Yes RB
NO DT

around
Could "around" be replaced with "approximately"?
Yes RB
No IN

as
Does "as" have a meaning similar to "so" (e.g. “This one is not as good")?
Yes RB
No IN

back
Is "back" modifying a noun (e.g. "back door")?
Yes JJ
No start

both
Is "both" being used with "and" (e.g. "Both the girls and the boys can do it.")?
Yes CC
No DT

bottom

Is "bottom" modifying a noun (e.g. "bottom drawer")?
Yes JJ
No start

but
Is "but" used like...
"except" ("everybody but me") IN
"only" ("We can but try") RB
None of the above CC

coming
Is "coming" being used like "upcoming"?
Yes JJ
No VBG

data
"Data" is used as a singular or plural noun. Is data being used as a:
Singular noun ("The data is surprising") NN
Plural noun ("The data are surprising") NNS
It is unclear from the context NN|NNS

dear
Is "dear" being used like:
"Oh dear" or "Dear me"? UH
"Yes, dear"? NN
"Dear Bob"? JJ
Other usage start

down
Is "down" being used to refer to currency or commodity prices?
Yes RB
No prep-test

either
Is "either" being used with "or" (e.g. "Either the girls or the boys can do it.")?
Yes CC
No DT

far
Is "far" used like:
An Adverb (e.g. "She lives far away") RB
An Adjective (e.g. "The far end of the field") JJ

for

Is "for" being used like "because"?
Yes CC
No IN

front
Is "front" modifying a noun (e.g. "front door")?
Yes JJ
No start

her
Would "her" be replaced by "his" or "him"?
His PRP$
Him PRP

his
Would “his” be replaced by “her” or “hers”?
Her PRP$
Hers PRP

however
Is "however" being used like "although" or "nevertheless", and could be removed from the sentence? (e.g. "However, the plans did not work")
Yes RB
No WRB

less
Is “less" referring to:
less of an amount ("You should eat less food") JJR
less of an action ("You should eat less (frequently)") RBR
the subtraction operator ("Five less two is three") CC

over
Is “over" being used as mathematical operator?
Yes CC
No start

may
Is "may" a noun?
Yes noun-start
No MD

minus
Is “minus" being used as mathematical operator?
Yes CC

No start

more
Is “more":
Referring to more of an amount or object ("You should eat more food", "It grows to five feet or more", or "more of the same") JJR
Replaceable by an adverb (such as "almost"), referring to more of an action ("You should run more") or modifies an adverb (e.g. "more carefully") RBR

much
Is "much" an adjective or adverb?
Adjective (e.g. "He doesn't have much energy left.")JJ
Adverb (e.g. "That's much better.") RB

near
Is "near" a preposition, adjective, or adverb?
Preposition (e.g. "We were near the station.") IN
Adjective (e.g. "The near side of the moon.") JJ
Adverb (e.g. "Her record is near perfect.", "They had gotten quite near.") RB

neither
Is "neither" being used with "nor" (e.g. "Neither the girls nor the boys can do it.")?
Yes CC
No DT

next
Is "next" an adjective, adverb, or preposition (archaic usage)?
Adjective (e.g. "The next train") JJ
Adverb (e.g. "They live next to me") RB
Preposition (e.g. "I grasp the hands of those next me.") IN

no
Is "no":
In the same location where "the" or "a" would be used? (e.g. "there is no answer yet"-->"there is an answer yet") DT
Being used as the opposite of "yes"? UH
Other (e.g. "This is no longer an issue") RB

one
Can "one":
Be replaced by "he", "she", or "it"? (e.g. "One shouldn't try this at home") PRP
Be pluralized or modified by an adjective? (e.g. "The one who cares."-->"The (nice) ones who care.") NN
None of the above (e.g. "one of the reasons", "one dollar") CD

only
Can "only" be replaced by "sole"?(e.g. "the only solution"-->"the sole solution")
Yes JJ
No RB

other
Could "other" be pluralized? (e.g. "This one is good but the other is bad" or "This one is good but the others are bad")
Yes NN
No start

plus
Is “plus" being used as mathematical operator?
Yes CC
No start

side
Is "side" modifying a noun (e.g. "side door")?
Yes JJ
No start

so
Is "so" being used like "so that"? (e.g. "I opened the door so he could leave.")
Yes IN
No RB

sooner
Could "sooner" be preceded (or is preceded) by "even"? (e.g. "I wish we could get there (even) sooner.")
Yes RBR
No RB

that
Could "that" be replaced by:
"which" (e.g. "the car that can't start") WDT
"the" (e.g. "I want that car.") DT
None of the above (e.g."He thought that the decision was wrong.") IN

then
Is "then" being used like "former"? (e.g. "The then president traveled to England.")
Yes JJ
No RB

there
Does "there" refer to a location and can be replaced by an adverb?
Yes (e.g. "I want to go there."-->"I want to go quickly.") RB
No (e.g. "There was a loud noise.") EX

times
Is “times" being used as mathematical operator?
Yes CC
No start

top
Is "top" modifying a noun (e.g. "top drawer")?
Yes JJ
No start

up
Is "up" being used to refer to currency or commodity prices?
Yes RB
No prep-test

very
Is "very" being used like "mere", "sheer", or "real"? (e.g. "The very thought")
Yes JJ
No RB

vice
Is "vice" being used in the same context as "vice president" or "vice principal"?
Yes NN
No start

well
Is "well" being used as a noun (e.g. "The well was full of water"), the opposite of "sick" (e.g. "He is feeling well"), or another usage?
Noun NN
Opposite of "sick" JJ
Other usage RB

what
Does "what" immediately precede a noun (not including pronouns) and any adjectives it may have? (e.g. "What kind do you want?", "Tell me what book to buy.")
Yes WDT
No WP

whatever

Does "whatever" immediately precede a noun (not including pronouns) and any adjectives it may have? (e.g. "Whatever events happen, we will be alright.", "Sell whatever books you own.")
Yes WDT
No WP

when
Is "when" used to refer to time? (e.g. "I know when he left")
Yes WRB
NO IN

will
Is "will" a noun, a verb (e.g. "Tom willed him to leave."), or a modal verb (e.g. "Sarah will visit her relatives.")
Noun noun-start
Verb verb-start
Modal MD

worth
Does worth precede a value or quantity? (e.g. "worth ten dollars", "worth a lot" etc.)
Yes IN
No start

yet
Could "yet" be replaced by "but"? (e.g."I like this, yet I wouldn't eat it again")
Yes CC
No RB

Appendix C: Words Automatically Tagged by the ITG

The following appendix lists the words and punctuation automatically tagged by the ITG. Note that some post-processing may change the POS tag shown. Collocations are also shown.

how WRB
where WRB
why WRB
whence WRB
whereby WRB
wherein WRB
whereupon WRB
a DT
an DT
every DT
the DT
another DT
some DT
each DT
these DT
this DT
those DT
which WDT
whichever WDT
to TO
can MD
could MD
might MD
must MD
ought MD
shall MD
should MD
would MD
'd MD
'll MD
ca MD
not RB

n't	RB
's	POS
'	POS
who	WP
whom	WP
whose	WP$
I	PRP
you	PRP
he	PRP
him	PRP
she	PRP
it	PRP
we	PRP
us	PRP
they	PRP
mine	PRP
yours	PRP
hers	PRP
ours	PRP
theirs	PRP
myself	PRP
yourself	PRP
himself	PRP
herself	PRP
itself	PRP
ourselves	PRP
yourselves	PRP
themselves	PRP
them	PRP
my	PRP$
your	PRP$
its	PRP$
our	PRP$
their	PRP$
and	CC
nor	CC
or	CC
,	,
.	.
$	$
:	:
;	:
...	:
%	NN

`` ``
" "
? .
# #
-- :
((
))
due JJ
many JJ
most JJS
people NNS
plenty NN
rather RB
such JJ
minimum NN
maximum NN
been VBN
i.e. FW
e.g. FW

Collocations:

Question Marks represent tags that are not defined in the collocation.

at all IN DT
all but RB RB
all right RB JJ
another one DT NN
closer to RBR TO
close to RB TO
due to IN TO
each other DT JJ
far from RB ?
farther from RBR ?
further from RBR ?
to see fit TO ? JJ
that 's hers ? ? PRP
that 's his ? ? PRP
a little bit ? JJ ?
a little more ? JJ ?
a lot ? NN
that 's mine ? ? PRP

nearer to RBR TO
near to RB TO
well off ? RP
better off ? RP
badly off ? RP
worse off ? RP
that 's ours ? ? PRP
It 's all over ? ? ? RB
It 's over ? ? RB
rather than IN IN
all right RB JJ
see fit ? JJ
so as to IN IN TO
so as not to IN IN RB TO
so that IN IN
that 's theirs ? ? PRP
that 's yours ? ? PRP
will be MD VB
very near RB RB

Appendix D: Expert Annotated ("Gold Standard") Corpus

Originally/RB ,/, Debus/NNP signed/VBD with/IN the/DT MLB/NNP St./NNP Louis/NNP Cardinals/NNPS out/IN of/IN the/DT Northern/NNP League/NNP ,/, but/CC his/PRP$ contract/NN was/VBD waived/VBN after/IN never/RB making/VBG an/DT appearance/NN ./.

Morgan/NNP studied/VBD music/NN at/IN London/NNP 's/POS Trinity/NNP College/NNP of/IN Music/NNP and/CC began/VBD his/PRP$ career/NN in/IN concert/NN work/NN and/CC radio/NN ./.

In/IN September/NNP 1943/CD it/PRP was/VBD in/IN training/NN near/IN Rome/NNP and/CC fought/VBD the/DT Germans/NNPS as/IN part/NN of/IN the/DT Corpo/NNP d'Armata/NNP Motocorazzato/NNP before/IN surrendering/VBG ./.

Even/RB in/IN those/DT days/NNS ,/, fund-raising/NN was/VBD necessary/JJ and/CC the/DT Sisters/NNPS organized/VBD several/JJ concerts/NNS to/TO clear/VB a/DT 1,500.00/CD debt/NN ./.

Matt/NNP was/VBD 19/CD and/CC too/RB old/JJ to/TO qualify/VB for/IN help/NN from/IN the/DT Make-A-Wish/NNP Foundation/NNP ./.

The/DT author/NN considers/VBZ examples/NNS such/JJ as/IN Wikipedia/NNP and/CC MySpace/NNP in/IN his/PRP$ analysis/NN ./.

She/PRP joined/VBD the/DT Queens/NNP County/NNP District/NNP Attorney/NNP 's/POS Office/NNP in/IN 1974/CD ,/, where/WRB she/PRP headed/VBD the/DT new/JJ Special/NNP Victims/NNPS Bureau/NNP that/IN dealt/VBD with/IN sex/NN crimes/NNS ,/, child/NN abuse/NN ,/, and/CC domestic/JJ violence/NN ./.

The/DT years/NNS of/IN isolation/NN have/VBP driven/VBN him/PRP mad/JJ ,/, and/CC he/PRP seeks/VBZ distraction/NN in/IN the/DT playing/NN of/IN games/NNS ./.

Many/JJ firearms/NNS ,/, particularly/RB older/JJR firearms/NNS ,/, had/VBD a/DT notch/NN cut/VBN into/IN the/DT hammer/NN allowing/VBG half-cock/NN ,/, as/IN this/DT position/NN would/MD neither/CC allow/VB the/DT gun/NN to/TO fire/VB

nor/CC permit/VB the/DT hammer-mounted/JJ firing/NN pin/NN to/TO rest/VB on/IN a/DT live/JJ percussion/NN cap/NN or/CC cartridge/NN ./.

Several/JJ new/JJ USB/NNP devices/NNS (/(especially/RB high-speed/JJ wireless/JJ WAN/NNP stuff/NN ,/, there/EX seems/VBZ to/TO be/VB a/DT chipset/NN from/IN Qualcomm/NNP offering/VBG that/DT feature/NN)/) have/VBP their/PRP$ Microsoft/NNP Windows/NNP device/NN drivers/NNS onboard/JJ ;/: when/WRB plugged/VBN in/IN for/IN the/DT first/JJ time/NN they/PRP act/VBP like/IN a/DT USB/NNP flash/NN drive/NN and/CC start/VBP installing/VBG the/DT device/NN driver/NN from/IN there/RB ./.

Ending/NNP Aging/NNP describes/VBZ de/NNP Grey/NNP 's/POS proposal/NN for/IN eliminating/VBG aging/NN as/IN a/DT cause/NN of/IN debilitation/NN and/CC death/NN in/IN humans/NNS ,/, and/CC restoring/VBG the/DT body/NN to/TO an/DT indefinitely/RB youthful/JJ state/NN ,/, a/DT project/NN plan/NN that/WDT he/PRP calls/VBZ the/DT ``/`` Strategies/NNPS for/IN Engineered/NNP Negligible/NNP Senescence/NNP "/" ,/, or/CC ``/`` SENS/NNP "/" ./.

He/PRP debuted/VBD with/IN them/PRP in/IN 1996/CD ,/, and/CC he/PRP finished/VBD second/RB to/TO Todd/NNP Hollandsworth/NNP in/IN Rookie/NNP of/IN the/DT Year/NNP Award/NNP balloting/NN ./.

The/DT Dungeness/NNP River/NNP is/VBZ a/DT 28-mile/JJ (/(45/CD km/NN)/) long/RB river/NN located/VBN in/IN the/DT Olympic/NNP Peninsula/NNP in/IN the/DT U.S./NNP state/NN of/IN Washington/NNP ./.

Sheffield/NNP played/VBD Phileas/NNP Fogg/NNP III/NNP ,/, the/DT great/JJ grandson/NN of/IN Phileas/NNP Fogg/NNP ./.

This/DT was/VBD to/TO encourage/VB new/JJ manufacturers/NNS to/TO enter/VB the/DT series/NN ,/, with/IN a/DT more/RBR even/RB field/NN of/IN cars/NNS and/CC cheaper/JJR running/VBG costs/NNS ./.

The/DT first/JJ team/NN to/TO get/VB their/PRP$ key/NN and/CC solve/VB their/PRP$ puzzle/NN wins/VBZ the/DT challenge/NN ./.

Since/IN 1992/CD he/PRP has/VBZ worked/VBN as/IN a/DT professor/NN of/IN philosophy/NN at/IN Loyola/NNP University/NNP Chicago/NNP ,/, where/WRB he/PRP holds/VBZ appointments/NNS in/IN the/DT philosophy/NN department/NN and/CC the/DT Parmly/NNP Sensory/NNP Sciences/NNPS Institute/NNP ./.

The/DT current/JJ editor/NN is/VBZ Cristanne/NNP Miller/NNP (/(University/NNP at/IN Buffalo/NNP ,/, The/DT State/NNP University/NNP of/IN New/NNP York/NNP)/) ./.

Taiwan/NNP High/NNP Speed/NNP Rail/NNP (/(abbreviated/VBN to/TO THSR/NNP or/CC HSR/NNP)/) is/VBZ a/DT high-speed/JJ rail/NN line/NN that/WDT runs/VBZ approximately/RB 345/CD km/NNS (/(214/CD mi/NNS)/) along/IN the/DT west/NN coast/NN of/IN the/DT Republic/NNP of/IN China/NNP (/(Taiwan/NNP)/) from/IN the/DT national/JJ capital/NN of/IN Taipei/NNP to/TO the/DT southern/JJ city/NN of/IN Kaohsiung/NNP ./.

The/DT overall/JJ champions/NNS were/VBD Pirmin/NNP Zurbriggen/NNP and/CC Erika/NNP Hess/NNP ,/, both/DT of/IN Switzerland/NNP ./.

It/PRP is/VBZ endemic/JJ to/TO Venezuela/NNP ./.

Palais/NNP Coburg/NNP was/VBD designed/VBN in/IN 1839/CD by/IN architect/NN Karl/NNP Schleps/NNP in/IN Neoclassical/JJ style/NN ,/, and/CC built/VBN from/IN 1840/CD to/TO 1845/CD by/IN Prince/NNP Ferdinand/NNP of/IN Saxe-Coburg/NNP and/CC Gotha/NNP atop/IN the/DT Braunbastei/NNP (/(Brown/NNP Bastion/NNP)/) ,/, a/DT part/NN of/IN the/DT Vienna/NNP city/NN defences/NNS dating/VBG to/TO 1555/CD ./.

The/DT school/NN is/VBZ intended/VBN for/IN students/NNS of/IN Indian/JJ nationality/NN ,/, nevertheless/RB it/PRP does/VBZ have/VB some/DT students/NNS are/VBP from/IN Pakistan/NNP ,/, Egypt/NNP ,/, and/CC Sri/NNP Lanka/NNP too/RB ./.

The/DT current/JJ President/NNP is/VBZ Nicolas/NNP Sarkozy/NNP ,/, who/WP was/VBD elected/VBN in/IN the/DT 2007/CD election/NN ./.

The/DT line/NN is/VBZ part/NN of/IN the/DT Line/NNP 1/CD of/IN Trans-European/NNP Transport/NNP Networks/NNPS (/(TEN-T/NNP)/) ./.

Turnout/NN ,/, however/RB ,/, was/VBD only/RB 42/CD %/NN of/IN the/DT electorate/NN by/IN far/RB the/DT lowest/JJS in/IN any/DT election/NN since/IN the/DT restoration/NN of/IN democracy/NN in/IN the/DT 1970s/NNS ./.

Honey/NNP joined/VBD the/DT Royal/NNP Air/NNP Force/NNP in/IN 1961/CD ./.

He/PRP has/VBZ the/DT distinction/NN of/IN directing/VBG the/DT series/NN finales/NNS for/IN Star/NNP Trek/NNP :/: Deep/NNP Space/NNP Nine/NNP ,/, Star/NNP Trek/NNP :/: Voyager/NNP and/CC Star/NNP Trek/NNP :/: Enterprise/NNP ./.

He/PRP is/VBZ also/RB standing/VBG as/IN a/DT mayoral/JJ candidate/NN in/IN the/DT Auckland/NNP mayoral/JJ election/NN ,/, 2010/CD ./.

Belfast/NNP Botanic/NNP Gardens/NNPS opens/VBZ as/IN the/DT private/JJ Royal/NNP Belfast/NNP Botanical/NNP Gardens/NNPS ./.

Eric/NNP Abetz/NNP (/(born/VBN 25/CD January/NNP 1958/CD in/IN Stuttgart/NNP ,/, West/NNP Germany/NNP)/) ,/, has/VBZ been/VBN a/DT Liberal/NNP Party/NNP member/NN of/IN the/DT Australian/JJ Senate/NNP since/IN February/NNP 1994/CD ,/, representing/VBG the/DT state/NN of/IN Tasmania/NNP ./.

Dr./NNP Don/NNP Olive/NNP ,/, a/DT self-proclaimed/JJ nuclear/JJ physicist/NN and/CC professor/NN of/IN science/NN at/IN the/DT university/NN maintains/VBZ and/CC opens/VBZ the/DT observatory/NN to/TO the/DT campus/NN and/CC community/NN on/IN occasion/NN ./.

By/IN the/DT mid-50s/NNS he/PRP was/VBD a/DT front/JJ running/JJ sports/NNS car/NN driver/NN ./.

The/DT Leopard/NNP Man/NNP 's/POS Story/NNP is/VBZ a/DT short/JJ mystery/NN story/NN by/IN Jack/NNP London/NNP ./.

Paintsville-Prestonsburg/NNP Combs/NNP Field/NNP covers/VBZ an/DT area/NN of/IN 25/CD acres/NNS (/(10/CD ha/NNS)/) at/IN an/DT elevation/NN of/IN 624/CD feet/NNS (/(190/CD m/NNS)/) above/IN mean/JJ sea/NN level/NN ./.

Plain-capped/NNP Ground-tyrant/NNP (/(M./NNP griseus/NN)/) was/VBD formerly/RB considered/VBN to/TO be/VB a/DT subspecies/NN of/IN M./NNP alpinus/NN but/CC is/VBZ now/RB commonly/RB treated/VBN as/IN a/DT separate/JJ species/NN ./.

Since/IN about/RB 1999/CD ,/, Crayfish/NNP Creek/NNP has/VBZ been/VBN subject/JJ to/TO heavy/JJ industrial/JJ logging/NN in/IN the/DT upper/JJ catchment/NN with/IN local/JJ residents/NNS attributing/VBG this/DT as/IN a/DT cause/NN for/IN a/DT significant/JJ loss/NN of/IN water/NN volume/NN ./.

The/DT notation/NN was/VBD introduced/VBN by/IN Adrien-Marie/NNP Legendre/NNP and/CC gained/VBD general/JJ acceptance/NN after/IN its/PRP$ reintroduction/NN by/IN Carl/NNP Gustav/NNP Jacob/NNP Jacobi/NNP ./.

Its/PRP$ main/JJ purpose/NN is/VBZ to/TO provide/VB a/DT complete/JJ Linux/NNP desktop/NN with/IN many/JJ popular/JJ applications/NNS and/CC tools/NNS ,/, yet/RB remain/VB small/JJ and/CC simple/JJ to/TO operate/VB ./.

Apart/RB from/IN military/JJ use/NN ,/, it/PRP was/VBD sold/VBN into/IN civilian/JJ use/NN ./.

Platymetopsis/NNP overali/NNP is/VBZ a/DT species/NN of/IN beetle/NN in/IN the/DT family/NN Carabidae/NNP ,/, the/DT only/JJ species/NN in/IN the/DT genus/NN Platymetopsis/NNP ./.

Paragraphs/NNS consist/VBP of/IN one/CD or/CC more/JJR sentences/NNS ./.

Stathmonotus/NNP culebrai/NN ,/, known/VBN commonly/RB as/IN the/DT Panamanian/JJ worm/NN blenny/NN in/IN the/DT United/NNP Kingdom/NNP ,/, is/VBZ a/DT species/NN of/IN chaenopsid/NN blenny/NN in/IN the/DT genus/NN Stathmonotus/NNP ./.

It/PRP is/VBZ divided/VBN into/IN two/CD sub-units/NNS :/: the/DT Danubian/NNP Flat/NNP in/IN the/DT south-west/NN ,/, with/IN eastern/JJ part/NN of/IN the/DT Zitny/NNP ostrov/NN island/NN ,/, and/CC the/DT Danubian/NNP Hills/NNP in/IN the/DT north/NN ,/, center/NN and/CC east/NN ./.

The/DT game/NN is/VBZ played/VBN by/IN two/CD people/NNS ,/, one/CD hold/VBP an/DT egg/NN vertically/RB another/DT tapping/VBG from/IN top/NN ./.

It/PRP grew/VBD over/IN the/DT next/JJ few/JJ years/NNS to/TO become/VB one/CD of/IN the/DT first/JJ really/RB large/JJ BBS/NNP systems/NNS ,/, which/WDT allowed/VBD its/PRP$ users/NNS to/TO carry/VB on/IN conversations/NNS with/IN thousands/NNS of/IN local/JJ residents/NNS ./.

Thokur-62/NNP is/VBZ a/DT census/NN town/NN in/IN Dakshina/NNP Kannada/NNP district/NN in/IN the/DT Indian/JJ state/NN of/IN Karnataka/NNP ./.

The/DT project/NN would/MD be/VB part/NN of/IN Norwegian/NNP County/NNP Road/NNP 585/CD and/CC financing/NN has/VBZ been/VBN secured/VBN through/IN the/DT Bergen/NNP Program/NNP ./.

The/DT host/NN country/NN ,/, Egypt/NNP ,/, achieved/VBD six/CD gold/NN medals/NNS but/CC also/RB shared/VBD the/DT joint/JJ highest/JJS total/JJ medal/NN count/NN with/IN Morocco/NNP ./.

Moves/NNS are/VBP communicated/VBN via/IN a/DT recognized/VBN chess/NN notation/NN ./.