A Test Collection for Part-of-Speech Tagging and Word Sense Disambiguation

Robert Krovetz
Lexical Research
Hillsborough, NJ 08844
rkrovetz@lexicalresearch.com

Abstract

We evaluate a focused test collection at the intersection of part-of-speech tagging and word-sense disambiguation. The collection targets words such as *train*, *novel*, and *lean*, where part-of-speech contrasts align with clear meaning differences. We use it to detect regressions across tagger versions, track quantitative and qualitative progress over time, and test robustness to orthographic variation. Experiments with the Stanford and TnT taggers show 68% accuracy, compared with 92% for a recent spaCy transformer model. Earlier taggers erred mainly on noun–verb distinctions; spaCy's errors more often involve noun–adjective distinctions. Uppercase text roughly doubles error rates for all taggers. We discuss common problems and propose directions for future testing.

Keywords: Evaluation, Part of Speech Tagging, Word Sense Disambiguation

1. Introduction

Part-of-Speech (POS) tagging is an important part of the NLP pipeline. It is done at an early stage, and errors can impact later stages of processing such as parsing, semantic role labeling, and word-sense disambiguation (WSD).

Our main interest is in the relationship between POS tagging and WSD. Many words, such as *train*, *novel*, and *lean*, show a correspondence between part-of-speech contrasts and clear differences in meaning. For example, *train* as a noun refers to a vehicle, whereas *train* as a verb means to instruct. A test collection focusing on such cases allows us to explore this intersection directly.

To that end, we constructed a focused lexical-sample test collection for POS tagging. This type of collection complements evaluation on running text and parallels the lexical-sample vs. all-words distinction used in the Senseval and SemEval competitions for WSD (Palmer et al. 2001; Pradhan et al. 2007).

At the same time, a lexical-sample approach also addresses several limitations of corpus-based accuracy measures. Because vocabulary follows a Zipfian distribution, many of the word types we would like to evaluate are rare, and errors involving them are easily missed. Moreover, conventional corpus-based evaluations include many tokens that belong to only one part of speech (e.g., *television*) or that are closed-class words (e.g., *the, and, for*). These dominate the accuracy score and obscure weaknesses in tagging rarer, open-class items. Mistakes on these words can have a high impact on downstream processing.

There are several aims for a test collection of this kind:

1. identifying cases where a new version of a system "breaks" something that previously worked.
2. measuring progress and identifying areas that still need improvement.
3. assessing robustness to contextual or orthographic variation.

The collection includes 500 words that each have at least two distinct senses and corresponding sentences illustrating each usage. Each instance is tagged with the correct part of speech. A sample of the words and contexts is shown in Table 1.

We used the test collection for three experiments:

1. comparing three versions of the spaCy tagger to identify regressions across versions.
2. comparing these with two early taggers, Stanford (Toutanova et al. 2003) and TnT (Brants 2000), to evaluate historical progress.
3. testing the robustness of tagging to orthographic variation by converting input sentences to uppercase.

The main contribution of the paper is the test collection itself. While WSD is usually applied only within a single part of speech, we extend the idea of homonymy (different word meanings) to distinctions across parts of speech. A secondary contribution is the application of a software-engineering concept, test collections, to the systematic evaluation of tagging systems.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11813–11821
11–16 May 2026.

Word	Tag	Label	Context
Abandon	VB	GIVE-UP	She will never abandon her child
Abandon	NN	UNRESTRAINED	She danced with abandon
Appropriate	JJ	SUITABLE	Is this an appropriate outfit
Appropriate	VB	STEAL	We will appropriate the jewels
Lean	VB	INCLINE	Please lean that against the wall
Lean	JJ	LOW-FAT	I would like some lean meat
Novel	JJ	NEW	That is a novel idea
Novel	NN	BOOK	I read a novel yesterday
Train	NN	LOCOMOTIVE	I took the train to work
Train	VB	EDUCATE	I will train you for the job
Accomplished	VBD	ACHIEVE	He accomplished what he wanted to do
Accomplished	JJ	SKILLED	She is an accomplished pianist
Biting	VBG	BITE	Try biting into this apple
Biting	JJ	CRUEL-HURTFUL	That was a biting remark
Defects	NNS	IMPERFECTION	They recalled the car because of defects
Defects	VBZ	POLITICS	What will happen if he defects to the USA

Table 1: Words in which a difference in part-of-speech is necessarily associated with a difference in meaning. The second column indicates the part of speech tag. The third column is a mnemonic that indicates at least one of the senses for each part-of-speech class. The fourth column is a context used for evaluation. The second part of the table shows ambiguity involving morphological variants.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the design of the test collection. Section 4 presents the tagging results and compares system performance. Section 5 evaluates the best-performing tagger (a transformer model) on tokens drawn from the British National Corpus. Section 6 presents an initial assessment using the Wikipedia. Section 7 discusses methodological implications and outlines plans for future work.

2. Related Work

Sparck Jones and Galliers (1996) provide an overview of evaluation in NLP. Lehmann et al. (1996) discuss the design of test suites, and the use of one such suite for parser evaluation. Manning (2011) and Macklovitch (1992) both examine problems associated with part-of-speech tagging. Dickinson and Meurers (2003) address issues with the gold standards used for evaluation and propose methods for automatically detecting tagging errors using n-grams. Chiche and Yitagesu (2022) provide a review of work on deep learning and part-of-speech tagging.

Sparck Jones and Galliers (1996, p. 168) define *test collections* as follows:

> "We mean ... by *test collections* material subsuming both input data and required output data or *answers* constituting *reference* data for evaluation". They further emphasize: "it is essential, in relation to what NLP is primarily about, to recognize the importance of test collections which include explicit answer data defining, for each input, what the system ought to give as output".

Test suites and *test collections* are concepts originating in Software Engineering and Information Retrieval. These ideas are complementary to the Machine Learning practice of dividing data into training, development, and test sets. Both approaches are concerned with system performance, but a test partition is intended to show how well a system has learned, rather than diagnose problems.

Mayhew et al. (2019) describe part-of-speech tagging and named-entity recognition when all text is in lowercase. They report the best results when training a language model on a mixture of cased and lower-cased text. In our part-of-speech dataset, only four test cases involved proper nouns affected by lowercasing. These were generally mis-tagged by earlier taggers, whereas a modern transformer-based tagger correctly labeled three of them even in lowercase. In contrast, uppercase text was often tagged incorrectly across all taggers.

3. Design of the Test for Part-of-Speech Tagging

The test has 500 words, and each word has at least two tags of interest. The tags of interest are based on semantic distinctions. For example, the difference between *train* as a noun or a verb, or

novel as a noun and an adjective. The test also includes examples of verbs and adjectival participles: *He is acting on a tip* vs. *He is the acting superintendent*, or *He accomplished what he wanted to do* vs. *She is an accomplished pianist*.

The distinctions across part-of-speech are not always *sufficient* conditions to identify a word's meaning. For example, *train* can also be used to refer to a *bridal train*, or to a *train of thought*. But the words in the test are intended to be cases where identifying part-of-speech is a *necessary* condition.

The test was created manually. Each sentence was designed to provide a context in which the intended part-of-speech of the target word is clear. It includes a few cases that were designed to be difficult, such as *minutes* as a plural noun ("How many minutes are left before we go?"), and as a singular noun ("The minutes of the meeting were very detailed"), and the same distinction for *premises*.[1] We are in the process of expanding the test using additional candidates from WordNet 3.0.[2] We wrote a script to find all synsets that differ in part-of-speech, and determined the overlap between the open-class words in the definitions. If there is no overlap in any of the definitions across part-of-speech, the word and parts-of-speech are proposed as a candidate.

We used the Stanford tagger (version 3.0), the TnT tagger (version 2.2c), and the tagger included in spaCy in evaluating the test. The Stanford tagger uses maximum-entropy, and the TnT tagger uses second order Markov models. The Stanford tagger was used with the left3words-wsj-0-18 model, the TnT tagger was used with wsj.tnt, and version 2.2.3 of spaCy was used with the en_core_web_sm model. We also compared performance against version 3.7.5 of spaCy using a transformer model.

We converted all of the example sentences to uppercase, and tagged them again. We compared the accuracy rate of the uppercase version against the original. We also converted all of the sentences to lowercase and again compared the results against the original.

We find text that is entirely in uppercase in old documents[3] and text that is entirely in lower text is common in tweets.

We report initial results of varying context. In particular, we note that taggers can sometimes be sensitive to punctuation differences. We therefore also looked at differences in the results when we varied the end of sentence punctuation between a period, a question mark, and an exclamation point. We also assessed contexts in which we put single quotation marks around each example sentence, as well as double quotations marks, and compared the tagging results against those using the original contexts.

4. Results

There were 1012 sentences that were used in the evaluation. In most cases an exact match was used between the expected tag and the observed tag. There was one exception to this: a VBP tag was allowed to match VB. This difference had no bearing on the semantic distinctions we were looking for.[4]

For the Stanford Tagger, there were 690 sentences that were correctly tagged. For the TnT tagger there were 686 that were correctly tagged. This means an overall accuracy of 68%.

We found mistakes that were in common between the taggers as well as cases where the taggers disagreed on an incorrect tag. For example, both the Stanford and TnT tagger were incorrect on the sentence *The fast went on for 20 days*, in which *fast* was tagged as an adjective by Stanford and as an adverb by TnT. In another sentence, *He is the acting principal of the school*, the word *acting* was incorrectly tagged as a verb form by both taggers.

The results with spaCy were much better than the other taggers. We found that 867 of the examples were correctly tagged in version 0.101, and 887 in version 2.2.3, for an accuracy rate of 86% and 88%, respectively. We will first discuss results with these two versions and how they compare with the Stanford and TnT taggers. We will then discuss how performance changed using the transformer model.

The evaluation illustrated several types of problems. The tags given below are for tags that were assigned by spaCy v0.101, but there were similar problems with all of the taggers. Those tags that were improved in spaCy 2.2.3 are indicated with an asterisk.

- **Inconsistent treatment of interjections**

  Boy! What a day is tagged correctly (interjection)
  Brother! What a day is tagged incorrectly (noun)

[1] The singular nouns were incorrectly tagged by all of the taggers.

[2] WordNet 3.0 is available from: https://wordnet.princeton.edu.

[3] For example, medical documents from Medline: https://pubmed.ncbi.nlm.nih.gov. Uppercase text also occurs in legal documents: https://public.resource.org/justice.gov/index.html

[4] VBP is the tag for non-3rd person singular present, e.g., "I go to work every day"; VB is the tag for the base form of the verb, e.g., "He wants to go now".

Hip! Hip! Hurray! is tagged correctly as interjections (UH). However, given the sentence surrounded by single quotes, it tags it as: interjection, interjection, proper noun. Given the sentence surrounded by double quotes, it tags it as: noun, interjection, proper noun. In version 2.2.3, all of the versions are tagged as proper noun, proper noun, interjection.

- **British English/American English differences**

  This is a cozy apartment (JJ) (correct)
  This is a cosy apartment (JJ) (correct)
  She put a cosy on the teapot (NN) (correct)
  She put a cozy on the teapot (JJ) (incorrect)*

- **Tagging adjectives as proper nouns**

  This is a Welsh painter (NNP) (incorrect)
  He is a Polish soldier (JJ) (correct)

- **Problems with adjectival participles**

  He is acting on a tip (VBG) (correct)
  He is the acting principal (VBG) (incorrect)
  What a cutting remark (VBG) (incorrect)

- **Tagging of noun/adjective ambiguity**

  He has a bum knee (JJ) (correct)[5]
  He is nothing but a bum (NN) (correct)

  He is a Hitchcock buff (NN) (correct)
  He has a buff piece of paper (NN) (incorrect)*

  This is a canary legal pad (JJ) (correct)
  He bought a canary (JJ) (incorrect)*

  The current is too strong (JJ) (incorrect)*
  What is the current rate (JJ) (correct)

- **Tagging of adjective/adverb ambiguity**

  It is cold outside (JJ) (correct)
  She knows the material cold (NN) (incorrect)

- **Surprises**

  Deck the halls with boughs of holly (correctly tags *Deck* as a verb). However, when the expression appears in double quotes, *Deck* is incorrectly tagged as a preposition. In version 2.2.3, *Deck* is tagged as a verb regardless of the quotation marks.

Converting the sentences to uppercase approximately doubles the error rate for the Stanford and TnT taggers, increasing from 32% to 65% and 62%, respectively.

[5]The tagging was incorrect in version 2.2.3.

Sometimes the Stanford and TnT taggers produced opposite results on upper-case text. For example, in lower-case text both taggers correctly tagged *grave* as an adjective in *This is a grave concern*, and as a noun in *They buried him in a grave*. In upper-case text, however, the Stanford tagger correctly tagged the adjective, but mistagged the noun, whereas the TnT tagger correctly tagged the noun, but mistagged the adjective.

Table 2 shows a sample of the changes in tagging between the early version of spaCy and version 2.2.3. The later version has better accuracy overall (88% versus 86%). But there are also cases where the new version does worse than the older one. The table shows cases that were improved, cases that were worse, and cases where both versions are incorrect.

We found that most frequent mistakes for both the Stanford and the TnT taggers were for tags involving nouns and verbs. In contrast, the most frequent mistakes for spaCy were for tags that involve nouns and adjectives. Of the remaining 77 errors on the test set, 65 (over 80%) involve the adjective tag, arising primarily from JJ/NN confusions, but also from JJ/VBG and JJ/RB ambiguities.

We found several changes in tagging as a result of varying punctuation. In addition to the previous problems with quotation marks, there were sentences that had different tags if the sentence ended in an exclamation point rather than a period. But the more general problem is the consistency of tagging despite variations in context. For example, in the sentence *We know that pretty little work has been done on the problem*, the word *pretty* is correctly tagged as RB (an adverb) by the Stanford and TnT taggers, and the first two versions of spaCy. But if the sentence was: *Pretty little work has been done on the problem*, then *Pretty* was tagged as NNP (proper noun) by the Stanford and TnT taggers, and as JJ (adjective) by the first version of spaCy. The second version of spaCy tags it correctly as RB.

We now turn to the results using the transformer model.

Deep learning has revolutionized computational linguistics. Version 3 of spaCy includes a transformer model, and we wanted to see how it compares with the earlier versions with respect to part-of-speech tagging. We used version 3.7.5 for our evaluation.

The transformer model has an accuracy of 92% on our part-of-speech dataset. This is an improvement from version 2.2.3 which was 88%. Like all of the taggers, performance decreases on upper-case text. The accuracy on an uppercase version of the sentences was 81%. That is, the accuracy rate is less on upper-case text than it is with the earlier version on text with normal case. In com-

Word	Expected	Example	v0.101	v2.2.3
broke	JJ	I am completely broke	JJ	VBN
bubbly	NN	Have a little bit of some bubbly	NN	RB
husky	NN	The husky and the cat got into a fight	NN	JJ
mighty	RB	This is a mighty strong case	RB	JJ
current	NN	The current is too strong	JJ	NN
invalid	NN	John is an invalid at the local hospital	JJ	NN
kind	JJ	She is a kind person	NN	JJ
minute	JJ	Put in a minute quantity of salt	NN	JJ
bubbly	JJ	She has a bubbly personality	NN	RB
cold	RB	She knows the material cold	NN	NN
fine	NN	What is the fine?	JJ	JJ
spoke	NN	The bicycle spoke was broken	VBD	VBD

Table 2: A comparison between two versions of spaCy. The first part of the table shows a sample of cases where accuracy was worse; the second part shows cases where accuracy was better; the third part shows cases where the tag was incorrect in both versions.

parision, the Stanford tagger had an accuracy rate of 68% on the dataset, but it was only 35% on the uppercase version. For the TnT tagger the accuracy rate was 38%.

We found cases in which an earlier version of spaCy sometimes outperformed the transformer model. For example, in *He took a moped to work*, the transformer tagged *moped* as an adjective, whereas spaCy 2.2.3 correctly tagged it as a noun. Tagging for this word is inconsistent: in *He went to work on a moped*, the transformer tags it as a noun. Interestingly, context beyond a single sentence can also influence tagging. In *He took a moped to work. He went to work on a moped.*, *moped* is tagged as a noun in both sentences.

There are also cases where the same error occurs in all versions of spaCy. For example, *acting* in *He is the acting principal of the school* is incorrectly tagged as a verb form, and *bats* in *John is completely bats* is tagged as a plural noun.

The en_core_web_sm model had comparable overall accuracy to the en_core_web_md and en_core_web_lg models. However, we found that the number of regressions for the larger models was roughly equal to the number of improvements.

The next two sections will discuss using the test collection in order to assess multiple tokens of each word+tag combination.

5. Assessment Using the British National Corpus

The British National Corpus (BNC) is a dataset of 100 million words that has been tagged using the CLAWS tagger (Burnard 2007), (Leech et al. 1994). The corpus is made up from diverse sources, and from both written and transcribed spoken text.[6] We used this corpus to evaluate both tagging accuracy and robustness. Accuracy was measured by comparing the transformer-based tagger's output with the tags provided in the BNC.[7] Robustness was tested by converting the sentences extracted from the BNC to uppercase, tagging them with the transformer-based tagger, and again comparing the tag with the one provided in the BNC.

All 500 words from our test collection were attested in the BNC, accounting for for 2.5 million tokens. However, not all part-of-speech combinations were observed for every word. For example, the word *abode* was attested only as a noun; the tensed verb form did not occur. Conversely, some words appeared with part-of-speech tags not included in the test collection. For example, *lean* is correctly tagged as a proper noun in the context "... directors like Carol Reed, David Lean, Michael Powell...".

In assessing token accuracy we did not want frequent words with many tokens for a particular part of speech to dominate the results. We therefore limited the assessment to ten tokens for each word+part-of-speech pair. This also excluded extremely rare cases that were attested fewer than ten times in the 100 million word corpus.

The resulting dataset included 466 word types and 11,990 sentences in which they occur. Of these 11,990 sentences, 4,560 (38%) contained multiple tags for the target word. For example, the following sentence had the word *abandon* tagged as both a noun and a verb: *Now, with reckless abandon, it promises to meddle with local-government structure.* The word was correctly tagged as a noun by the transformer, but we can-

[6] The corpus is available from: www.natcorp.ox.ac.uk

[7] We wrote a script to map between the CLAWS tagset and the one used in the Penn Treebank.

not assess whether that is correct since it is only one of the possibilities. In addition, we found that even with the ambiguous tags, sometimes both of them were incorrect. For example, for the sentence *"Get Mrs Long," one of the girls was ordered.*, the word *Long* was incorrectly indicated as being either an adverb or adjective.

We tested two hypotheses in the tagging assessment. The previous section showed that more than 80% of the remaining errors on the test involved distinctions between adjectives and other parts of speech. We therefore divided the attested word+part-of-speech pairs into two groups: those in which the word can function both as an adjective and as another part of speech (such as *novel* and *lean*), and those in which an adjectival interpretation is not possible (such as *train*). The first hypothesis is that the error rate will be higher in the first group than in the second group. The second hypothesis is that the error rate on the uppercase version of the extracted sentences will be higher than on the sentences with the original orthography. We limited the assessment to sentences in which the target word had only one part-of-speech tag. In addition, the evaluation focused on the word+part-of-speech pairs included in our test inventory. Tokens in which the BNC tag corresponded to a part of speech outside those pairs were excluded.

Both hypotheses were confirmed. Table 3 shows the results. There were 2952 tokens that involved adjectives as a possible part-of-speech for the target word, and 4310 tokens where an adjective was not a possibility.[8] The error rate (defined as a mismatch between the BNC tag and the tag assigned by the transformer) was 27.6% for the first group and 14.5% for the second group. For the uppercase version of the sentences, the error rates were 42.8% and 19.1%, respectively.

Because the BNC tags are not always correct, these rates should not be accepted as ground truth for tagging error. However, they reinforce the result that future assessments should focus on words that involve adjectives as a possible part-of-speech, and that performance can change for the worse when taggers are evaluated on upper-case text.

The median frequency of the assessed word+part-of-speech tags in the BNC was 925. This is why we used the word "rare" in the Introduction. In addition, the distribution is often skewed between one part-of-speech and another. All minority cases could be incorrectly tagged with only a slight effect on overall corpus-level accuracy. Indeed, the total number of tokens is only 2.5% of the 100 million word corpus. If all tokens for the 500 word types in the test collection were incorrectly tagged it would mean a drop of 2.5% in accuracy.

[8]We excluded 9 words that had more than two parts-of-speech in the test collection. In such cases the taggers could agree on a third tag outside the contrast being evaluated (e.g., *acting* tagged as NN rather than VBG).

6. Assessment Using the Wikipedia

The British National Corpus (BNC) is one of the largest balanced corpora with part-of-speech annotations. However, it was tagged using a tagger that is no longer state of the art, and it sometimes assigns multiple tags to a single word. These characteristics motivate the use of alternative resources for targeted evaluation. We therefore created a dataset that is released under a Creative Commons Attribution license. The dataset consists of sentences containing a target word from our test collection, with each sentence tagged using the spaCy tagger. The sentences were taken from a download of the English Wikipedia.

The dataset is provided as a file with the following columns: target word, silver tag, gold tag, tagged sentence, untagged sentence, and source. The target word is one of the 500 words in the test set, all of which are attested. The silver tag is assigned by spaCy 3.7.5 using the en_core_web_sm model, while the gold tag is currently a placeholder for manually verified annotations. Both tagged and untagged sentences are included to facilitate future comparisons. The source column records the title of the article in which the sentence appears.

We chose the en_core_web_sm model to generate silver tags in part because it is substantially faster than the transformer-based model. However, the comparison between the transformer model and a previous state-of-the-art tagger is also used to focus assessment. Cases where the systems disagree help identify contexts that should receive a gold tag, which can then be used to fine-tune current systems.

In the test collection, 76 words (77 sentences) were incorrectly tagged by the transformer model. From this set, we selected 15 words for further investigation at the token level. These words are listed in Table 4, broken down by the type of ambiguity. The words fall into three types of mistagging, with five words in each category: noun/adjective, adverb/adjective, and verb form/adjective.

For each word type we assessed up to 40 tokens: 20 where the transformer tag matched the silver tag, and 20 in which the tag differed. These were intended to include 10 tokens for each of the two word+part-of-speech combinations in the silver annotation. The number of evaluable tokens varied across groups (191, 164, and 196 tokens respectively out of a possible 200 tokens per ambiguity type). The groups did not yield sufficient tokens

	Count	Difference in Tagging	Uppercase Difference
Involves Adjective	2952	27.6%	42.8%
Does not involve Adjective	4310	14.5%	19.1%

Table 3: Tagging differences between a transformer-based tagger and the British National Corpus (BNC) tagging. The table shows the number of tokens that were assessed and the percentage that involved a difference in tagging. The results are broken down into those in which a word could be tagged as an adjective or not, and on uppercase versions of the sentences.

Tag Contrast	Words	Accuracy
NN/JJ	aerial, camp, crash, material,orange	78%/61%
RB/JJ	cold, dead, precious, real, short	66%/56%
VBG/JJ	acting, binding, pressing, prevailing, standing	78%/57%

Table 4: Tagging differences between a transformer-based tagger and a non-transformer spaCy tagger over sentences from the English Wikipedia. Words are grouped by the type of ambiguity. Percentages give average accuracy in agreement vs. disagreement subsets. Up to twenty tokens per word were evaluated for each condition.

for *crash*, *precious*, *short*, and *prevailing* as the taggers frequently agreed on incorrect tags for these words. As a result, there were too few instances in which the tags differed to meet the sampling target.

The table provides the average accuracy for each group by type of ambiguity. We report separate averages for cases where the taggers agree and where they disagree.

We found that accuracy is higher in cases where the silver tag and the transformer tag agree than in cases where they disagree. This is as we would expect. The silver tag serves as a useful prior. However, even within the subsets where the taggers agree, the accuracy rates show room for improvement.

Examining the changes in detail, we found several reasons for errors made by the transformer model:

1. Hyphenated contexts that include the target word. When the target word could be an adjective or an adverb, the transformer sometimes incorrectly tagged it as an adverb, as in *cold-hearted, real-valued, short-lived*. Similar issues arise with with other parts of speech: for example, *crash* was incorrectly tagged as an adjective in *non-crash*, *pre-crash*, and *post-crash*.

2. Certain types of ambiguity were more difficult than others. In particular, contrasts between adjectives and adverbs and between adjectives and verb forms were the hardest to tag.

3. Some errors were concentrated on particular words rather than particular contexts. For example, the word *prevailing* was often tagged as VBG even when used adjectivally.

We argue that it is unnecessary to evaluate every sentence in which a target word can function as an adjective. Part-of-speech distributions for ambiguous words are typically highly skewed. For example, *said* is overwhelmingly used as a tensed verb, but it occurs as an adjective in contexts such as *the said regulation*, or *the said district*. To focus evaluation efficiently, we identify the minority sense by counting how many sentences were tagged with the part-of-speech of interest. For example, if *lean* is tagged more often as an adjective, we evaluate the contexts in which it is tagged as a verb. As time and resources permit, a sample of the tagged sentences for the majority sense can also be evaluated. We are currently tagging minority sense instances with the transformer model to assess accuracy for other word types at the token level. As with the evaluation above, we will examine a sample of the cases where there is agreement with the silver standard, and cases where the tagging is different.

7. Discussion and Future Work

Extrinsic and intrinsic evaluations are both important. We believe there is a need for test collections that "stress test" different issues that are important for natural language processing.

Based on the errors observed in this study, we plan to pursue the following directions:

1. **Targeted test collections for frequent error types.** Many errors involved ambiguity between adjectives and other parts of speech. We will develop a test collection that focuses on such sources of error.

2. **Evaluation across diverse domains.** Performance is likely to vary across domains, particularly for low-frequency vocabulary and specialized contexts. We therefore plan to construct test collections drawn from multiple domains to assess generalizability.

3. **Interactions between part-of-speech tagging and named-entity recognition.** Tokens classified as PERSON, LOCATION, and ORGANIZATION, should generally be classified as proper nouns, yet named entities are often out-of-vocabulary terms. We will investigate how such cases affect tagging accuracy and develop evaluation sets that target these interactions.

4. **Error-driven evaluation using system comparisons.** Differences between the current state of the art and earlier systems can be used to identify informative test cases. This approach reduces annotation effort and supports the construction of focused gold-standards for model analysis and fine-tuning. However, it is not enough to just look at differences. More than 90% of the tokens were tagged identically, so it is also necessary to sample from cases where taggers are in agreement.

These observations highlight the importance of evaluation methodologies that go beyond overall accuracy and instead target specific sources of error and variability. As future work, we will extend our analysis to examine the impact of frequency and skew, defined as the ratio between the most frequent tag and the next most frequent tag.

8. Conclusion

This paper has described a test collection for part-of-speech tagging. The test is designed to provide a rich and detailed assessment of cases where a difference in part-of-speech is associated with semantic distinctions.

We evaluated the test against freely available taggers that are widely used in the community. We found variations between taggers in how the expressions are treated. Subsequent versions of the spaCy tagger sometimes resulted in lower accuracy on particular expressions, despite improvements in overall performance.

The Stanford and TnT taggers were similar in accuracy (68%) and in the types of errors they produced. The spaCy tagger was more accurate (92%), and made errors primarily on cases that are intuitively more difficult - such as distinguishing between nouns and adjectives, rather than nouns and verbs. In the test set, over 80% of the remaining tagging errors involve words that can function as an adjective.

Upper-case text was a significant problem. For all taggers the error rate approximately doubled when the text in the test collection was converted to uppercase.

We assessed tagging at the token level using the British National Corpus and the English Wikipedia. We conducted a focused evaluation of a transformer model on the Wikipedia and created a dataset for further evaluation.

The test collection, the tagged subset of the Wikipedia, and the evaluation data used in this study are publicly available under a Creative Commons Attribution 4.0 license (CC-BY-4.0). The files can be downloaded from https://github.com/rkrovetz/pos.

9. Bibliographical References

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA. Association for Computational Linguistics.

Lou Burnard. 2007. *Reference Guide for the British National Corpus (XML Edition)*. Oxford University Computing Services, Oxford, UK.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9:Article 10.

Markus Dickinson and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, volume 1, pages 622–628, Kyoto, Japan. Association for Computational Linguistics.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP - test suites for natural language processing. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Elliott Macklovitch. 1992. Where the tagger falters. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Montréal, Canada.

Christopher Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 171–189. Springer.

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6256–6261, Hong Kong, China. Association for Computational Linguistics.

Martha Palmer, Christiane Fellbaum, Scott Cotten, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of the SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Karen Sparck Jones and Julia Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*, volume 1083. Springer.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.