Using Amazon Mechanical Turk for Data Collection in Parts of Speech Tag Correction for Patent Claims

David Cinciruk

June 24, 2015

Table of Contents

Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

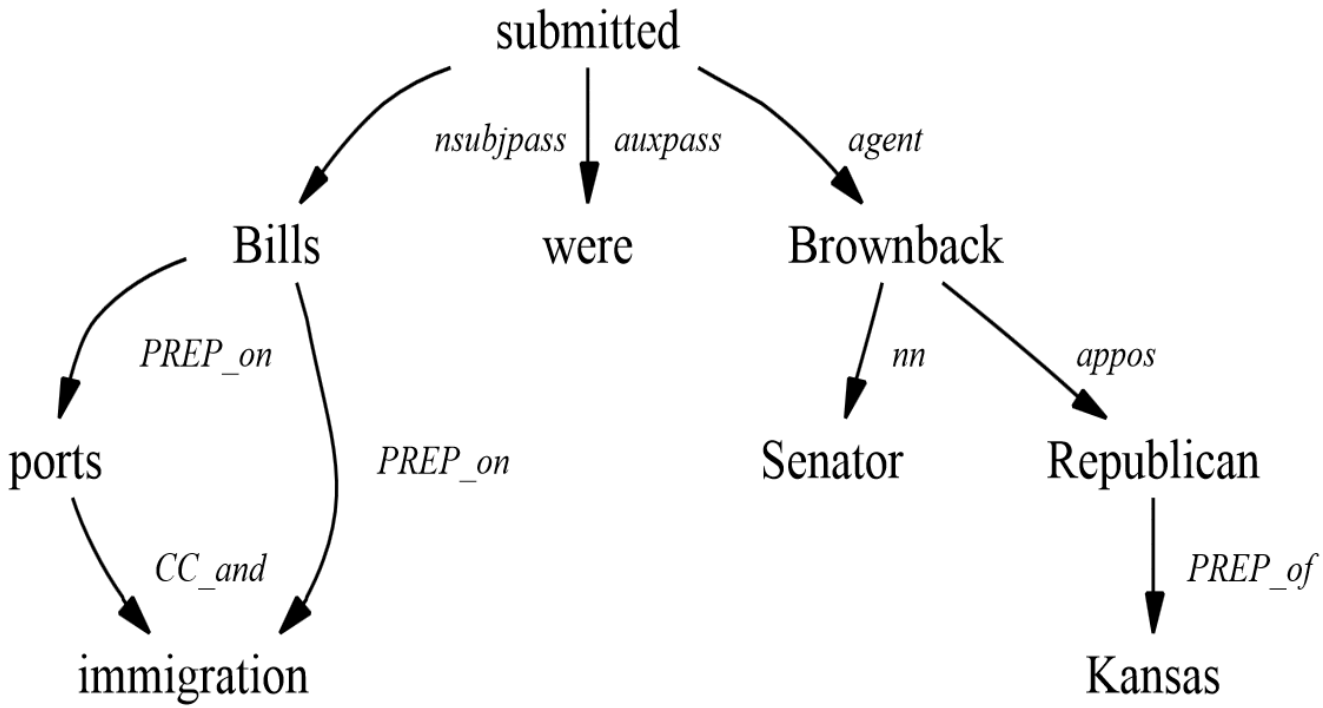
Automatic Correcter of Patent Claim POS Tags

Conclusion

Our Research

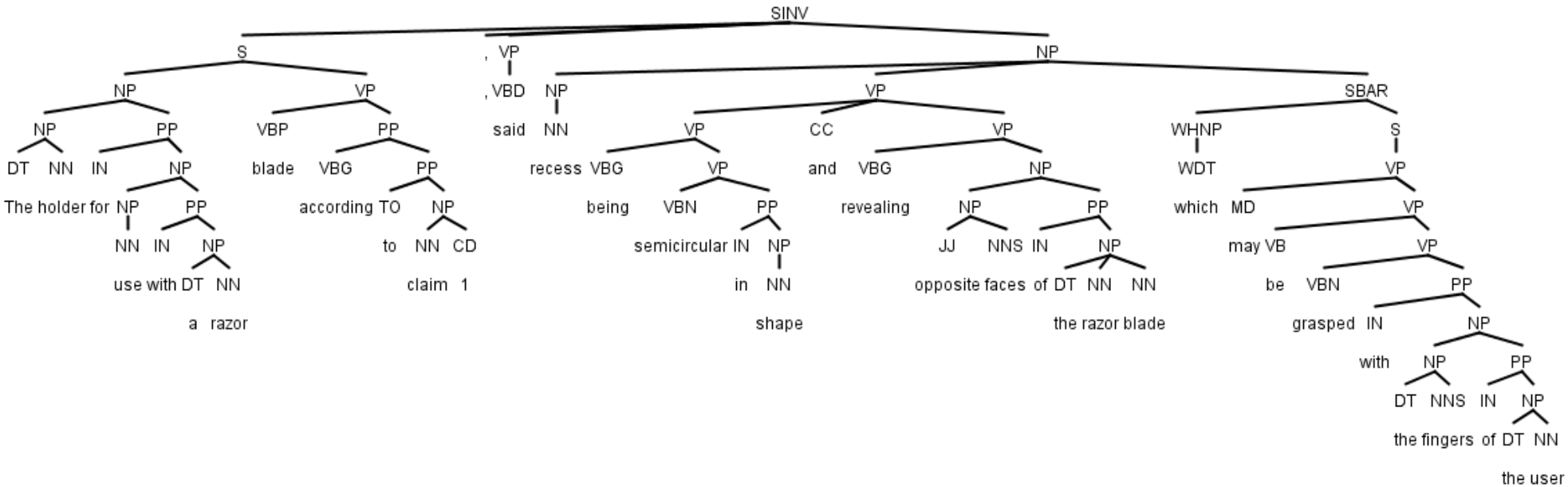
- Work is in Patent Processing
  - Mapping patents to other technical documents
  - Automated classification
  - Patent Retrieval
  - Patent Valuation
- Need to find a way to represent patents. NLP offers a solution with dependencies

Dependency Modeling

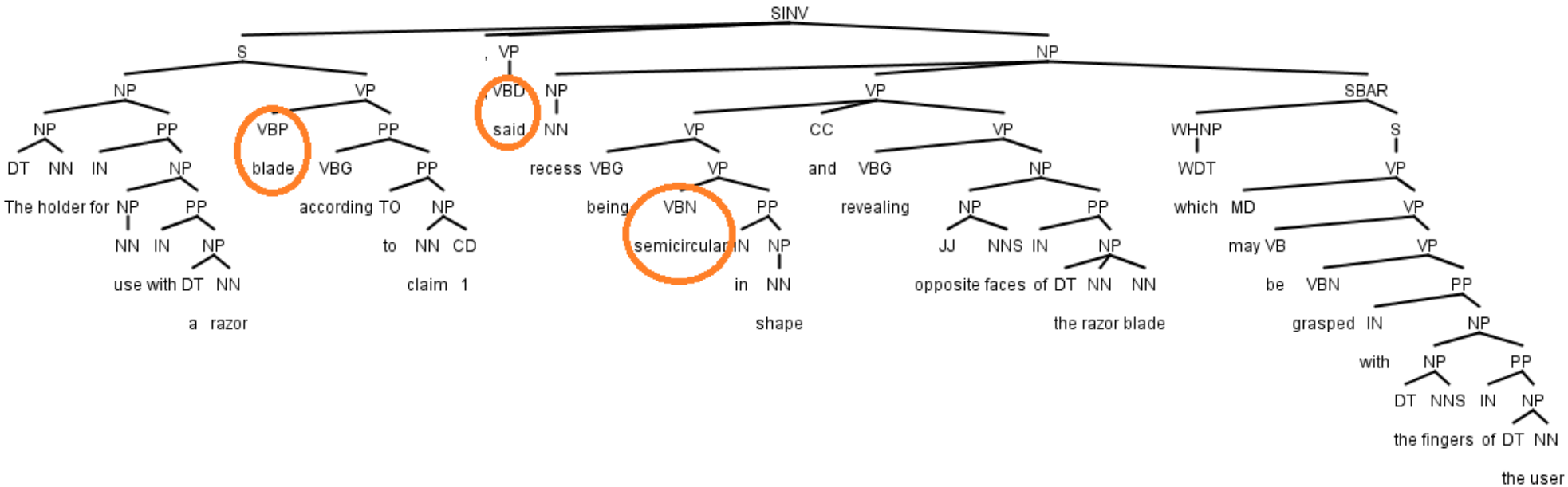


- Dependencies help to represent words by their relationships
- Formed by traversing the parse tree of a sentence or segment and noting two words and how they are linked together

NLP Parsers Do Not Work Well With Patent Claims!



Obvious Mislabeling of Words



Odd Language of Patent Claims

The holder for use with a razor blade according to claim 1, said recess being semicircular in shape and revealing opposite faces of the razor blade which may be grasped with the fingers of the user.

- Patents are the intersection of technical and legal speech
- Legally each patent claim must be one very long run-on sentence
- Features particular language whose meanings aren't standard meanings in normal English.
- Because of these, NLP software struggles to correctly tag patent claims

Trying to Correct the NLP

blade: VBP $\rightarrow$ NN
said: VBD $\rightarrow$ JJ
semicircular: VBN $\rightarrow$ JJ

- By forcing incorrect tags to a corrected tag, NLP softwares would be forced to reconstruct the parse tree and create dependencies more similar to true speech
- We need a large collection of hand-corrected patent claims in order to train a system to automatically correct patents
- Amazon Mechanical Turk allows us to gather this collection via crowdsourcing

Table of Contents

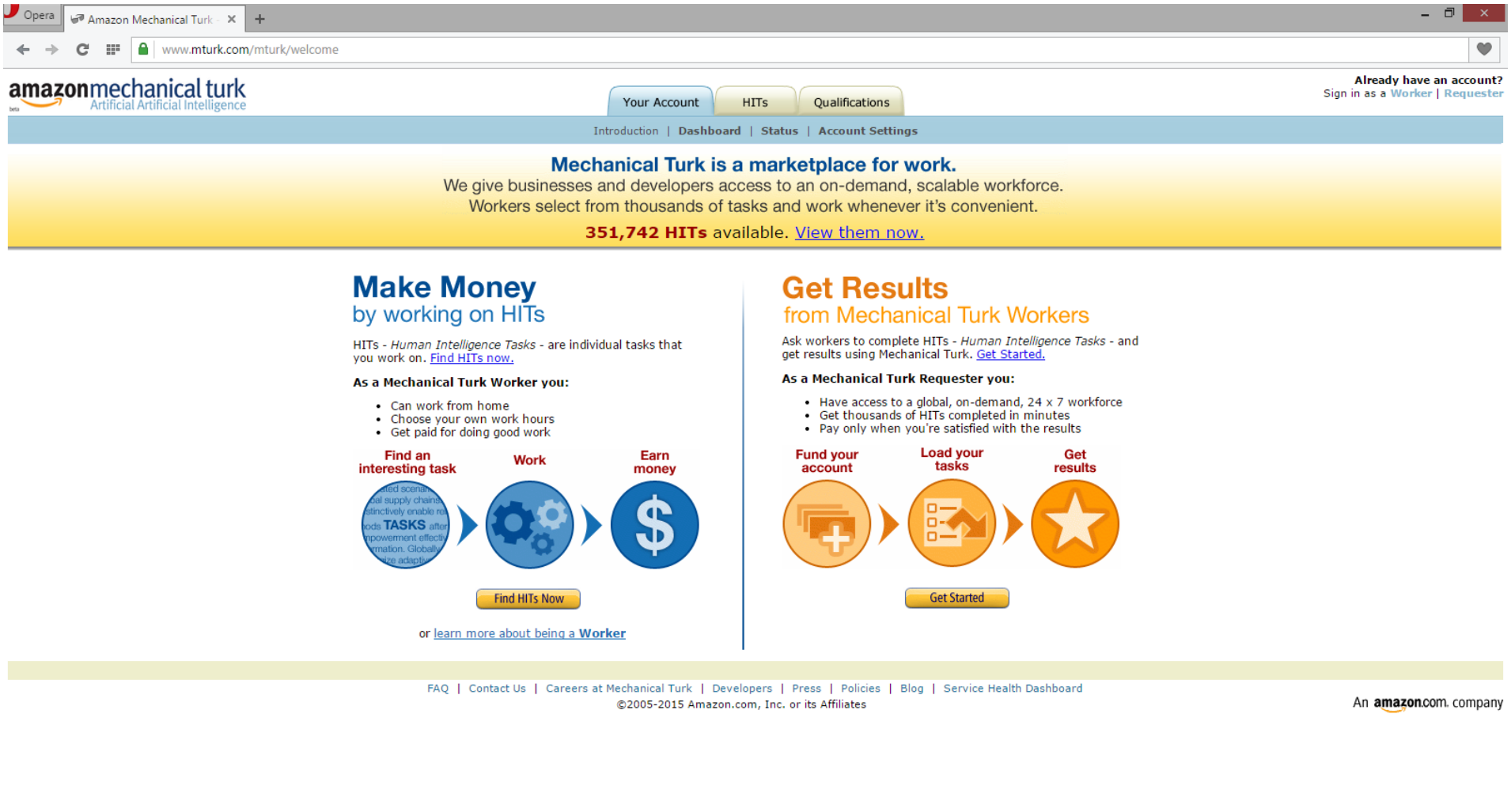
Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

Automatic Correcter of Patent Claim POS Tags

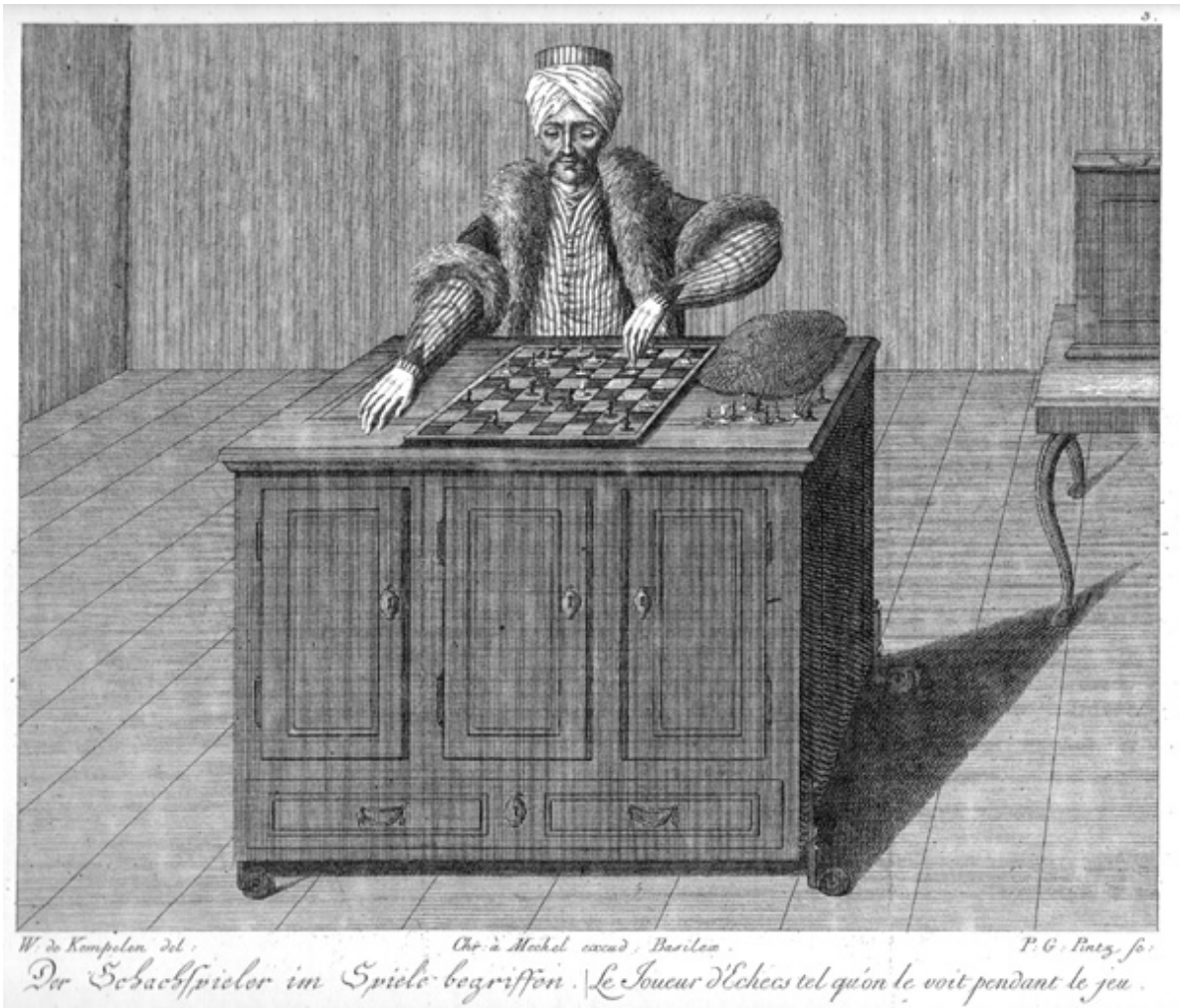
Conclusion

Artificial Artificial Intelligence



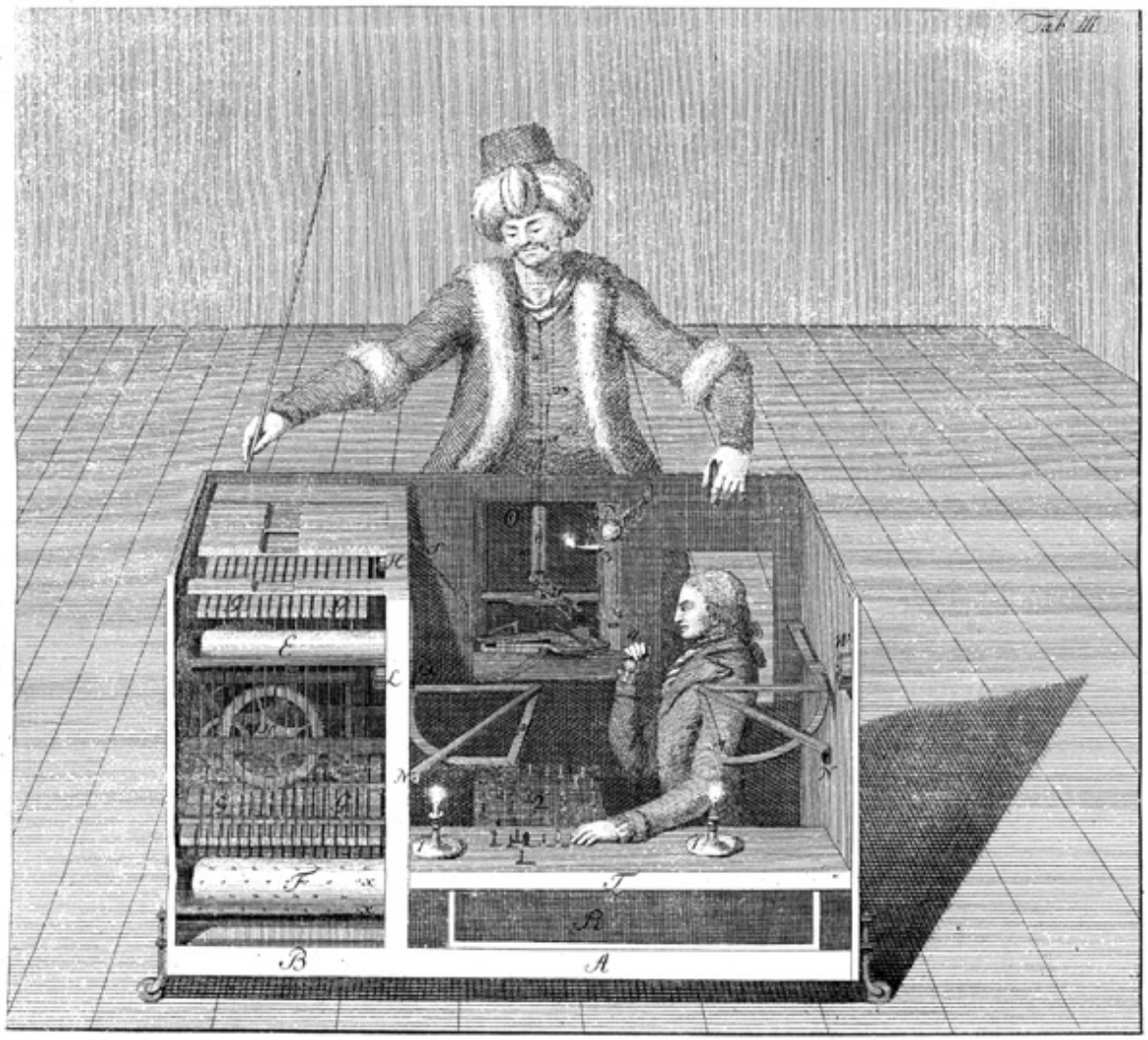
Crowdsourcing Internet Marketplace that enables individuals and organizations to coordinate the use of human intelligence to perform tasks that computers are unable to do.

The Turk



Origin comes from a supposed master Chess Playing Automaton built in the 1770 by Wolfgang von Kempelen

The Turk



Actually was a chess master hidden in the cabinet playing the game from underneath

Workers and Requesters

- **Worker**
- Also known as Turkers
- Performs tasks set up by requesters
- Work their own hours and choose their own tasks

- **Requester**
- Creates tasks for workers to do
- Can set qualifications on tasks to dislow unproven people from performing tasks.
- Can set up tests to determine if people are qualified to work on particular HITs or not

Our Use of Mechanical Turk

- We want to develop the most accurate dependencies for use in our future systems
- We have patent claims automatically segmented based off of semicolons and colons and POS tagged
- Tags may be incorrect - want to eventually be able to automatically correct the tags
- Need to gather lists of corrected tags
- Decided to use Mechanical Turk to gather corrected tags

Our Use of Mechanical Turk

- Want to create an easy to perform HIT that has high impact on our tagging.
- Most common problem that has a high impact on the dependencies are problems with words incorrectly tagged as verbs
- Chose the task of checking if words initially tagged as verbs are nouns or adjectives
- Turkers will label the verbs as noun, adjective, or verbs and assume all other tags are correct

The Setup

- **Initial:**
  - The/DT holder/NN for/IN use/NN with/IN a/DT razor/NN **blade/VBP according/VBG** to/TO claim/NN 1/CD ,/, **said/VBD** recess/NN **being/VBG semicircular/VBN** in/IN shape/NN and/CC **revealing/VBG** opposite/JJ faces/NNS of/IN the/DT razor/NN blade/NN which/WDT may/MD **be/VB grasped/VBN** with/IN the/DT fingers/NNS of/IN the/DT user/NN ./.
- **Corrected:**
  - The/DT holder/NN for/IN use/NN with/IN a/DT razor/NN **blade/NN according/VBG** to/TO claim/NN 1/CD ,/, **said/JJ** recess/NN **being/VBG semicircular/JJ** in/IN shape/NN and/CC **revealing/VBG** opposite/JJ faces/NNS of/IN the/DT razor/NN blade/NN which/WDT may/MD **be/VB grasped/VBN** with/IN the/DT fingers/NNS of/IN the/DT user/NN ./.

Table of Contents

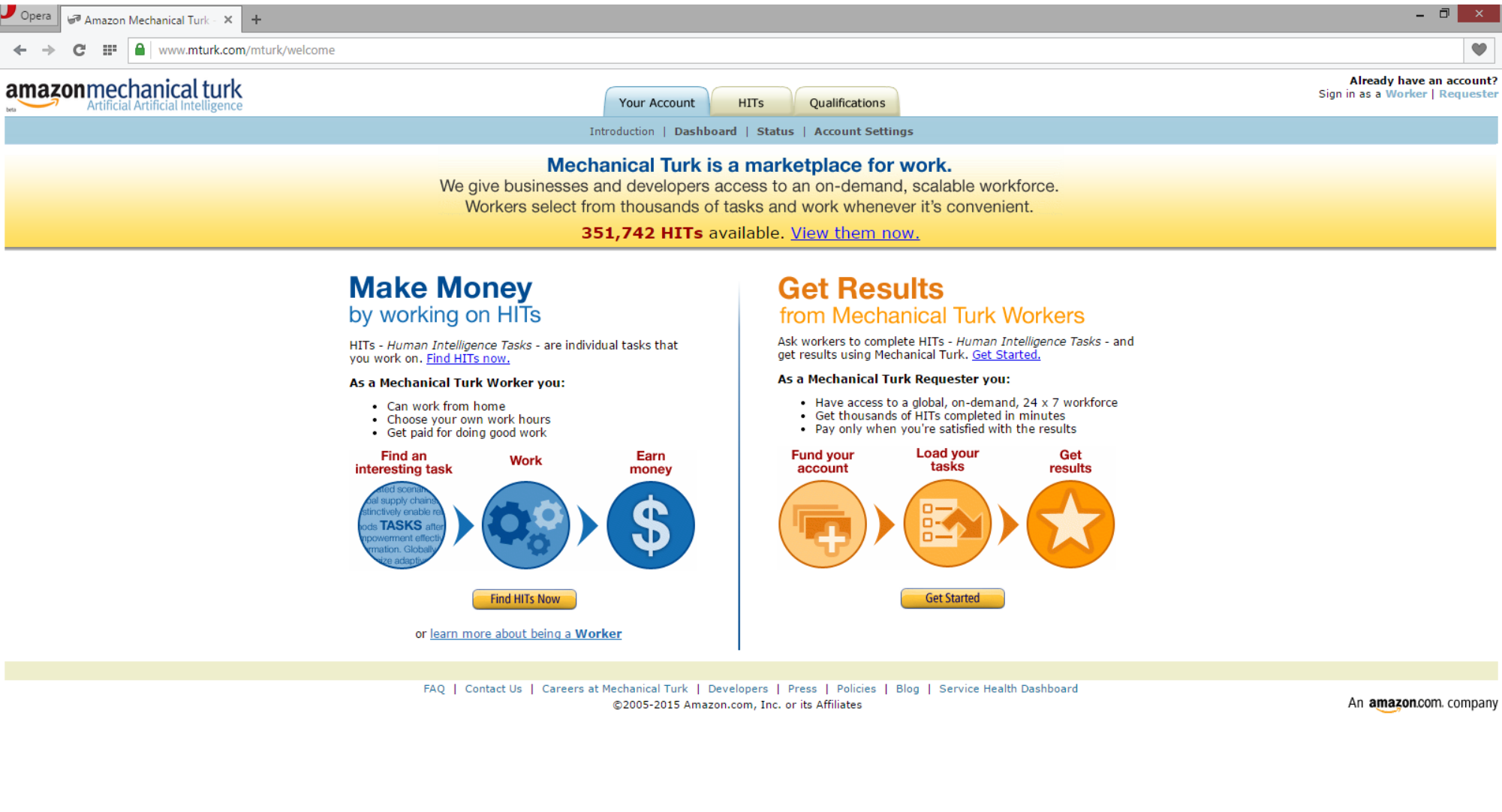
Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

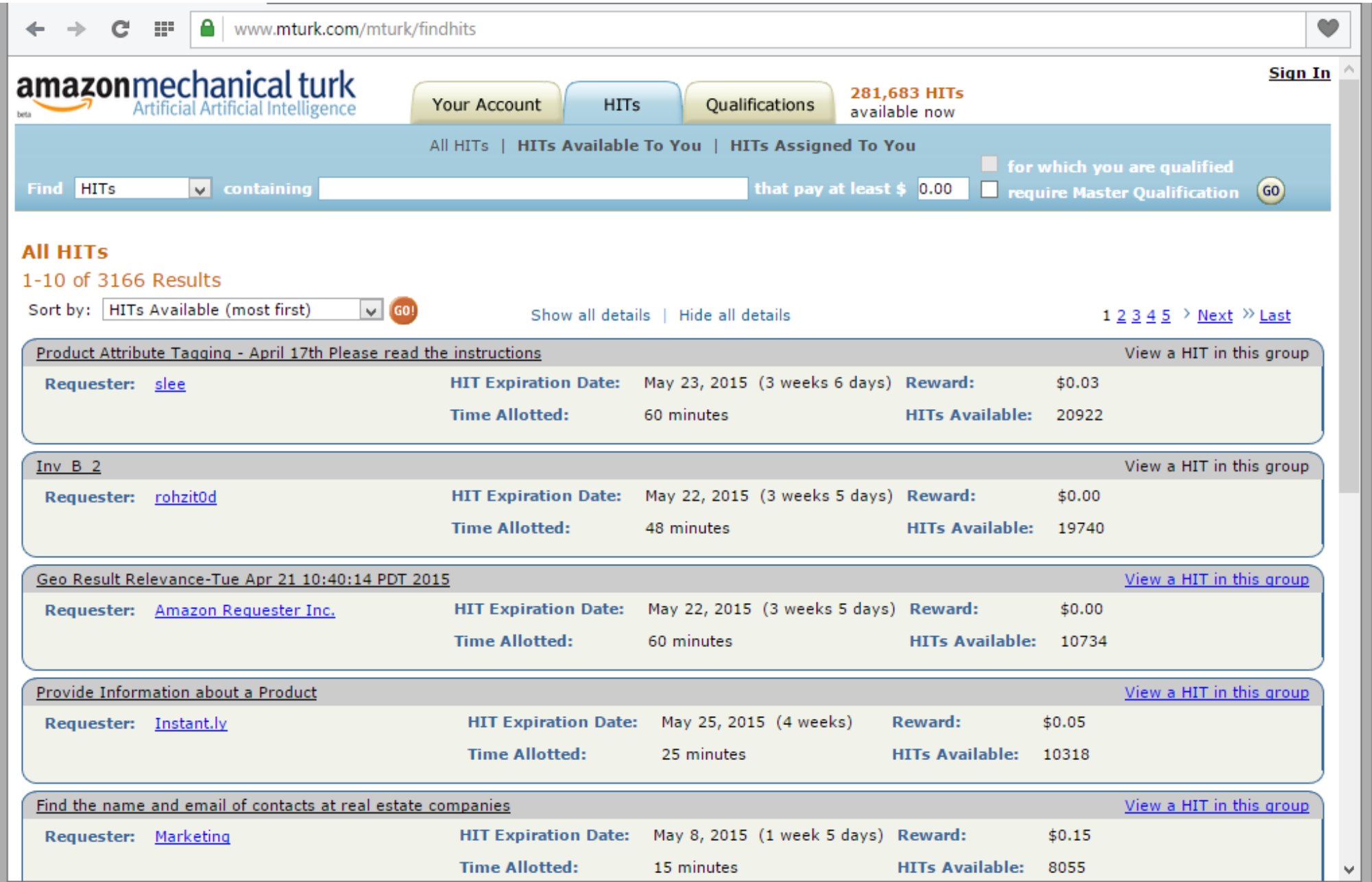
Automatic Correcter of Patent Claim POS Tags

Conclusion

Worker Homepage



HIT Search


www.mturk.com/mturk/findhits
amazonmechanical turk
Artificial Artificial Intelligence
Sign In
Your Account
HITs
Qualifications
281,683 HITs
available now
All HITs | HITs Available To You | HITs Assigned To You
Find HITs containing that pay at least $ 0.00
for which you are qualified
require Master Qualification
GO
All HITs
1-10 of 3166 Results
Sort by: HITs Available (most first) GO!
Show all details | Hide all details
1 2 3 4 5 › Next » Last
Product Attribute Tagging - April 17th Please read the instructions
View a HIT in this group
Requester: slee
HIT Expiration Date: May 23, 2015 (3 weeks 6 days)
Reward: $0.03
Time Allotted: 60 minutes
HITs Available: 20922
Inv_B_2
View a HIT in this group
Requester: rohzit0d
HIT Expiration Date: May 22, 2015 (3 weeks 5 days)
Reward: $0.00
Time Allotted: 48 minutes
HITs Available: 19740
Geo Result Relevance-Tue Apr 21 10:40:14 PDT 2015
View a HIT in this group
Requester: Amazon Requester Inc.
HIT Expiration Date: May 22, 2015 (3 weeks 5 days)
Reward: $0.00
Time Allotted: 60 minutes
HITs Available: 10734
Provide Information about a Product
View a HIT in this group
Requester: Instant.ly
HIT Expiration Date: May 25, 2015 (4 weeks)
Reward: $0.05
Time Allotted: 25 minutes
HITs Available: 10318
Find the name and email of contacts at real estate companies
View a HIT in this group
Requester: Marketing
HIT Expiration Date: May 8, 2015 (1 week 5 days)
Reward: $0.15
Time Allotted: 15 minutes
HITs Available: 8055

HIT Search

Type the text from the images, carefully. Productivity and bonuses guaranteed. — View a HIT in this group

Requester: CopyText Inc. **HIT Expiration Date:** Apr 27, 2015 (6 days 19 hours) **Reward:** $0.01

Time Allotted: 10 minutes **HITs Available:** 11652

Description: Type the text carefully

Keywords: ocr, copy text, bonus, easy

Qualifications Required:

HIT approval rate (%) is greater than 95

Transcribe up to 25 Seconds of Media to Text - Earn up to $0.12 per HIT! — View a HIT in this group

Requester: Crowdsurf Support **HIT Expiration Date:** Apr 18, 2016 (51 weeks 6 days) **Reward:** $0.08

Time Allotted: 15 minutes **HITs Available:** 11376

Description: Transcribe up to 25 seconds of media into text. Reward amount is variable based on media length. Additional reward amounts will be paid as a bonus.

Keywords: Transcribe, transcription, media, audio, English, type, typist, caption, subtitle

Qualifications Required:

Qualified to work on Transcription Tasks is not less than 900

HIT approval rate (%) is not less than 95

Our HIT

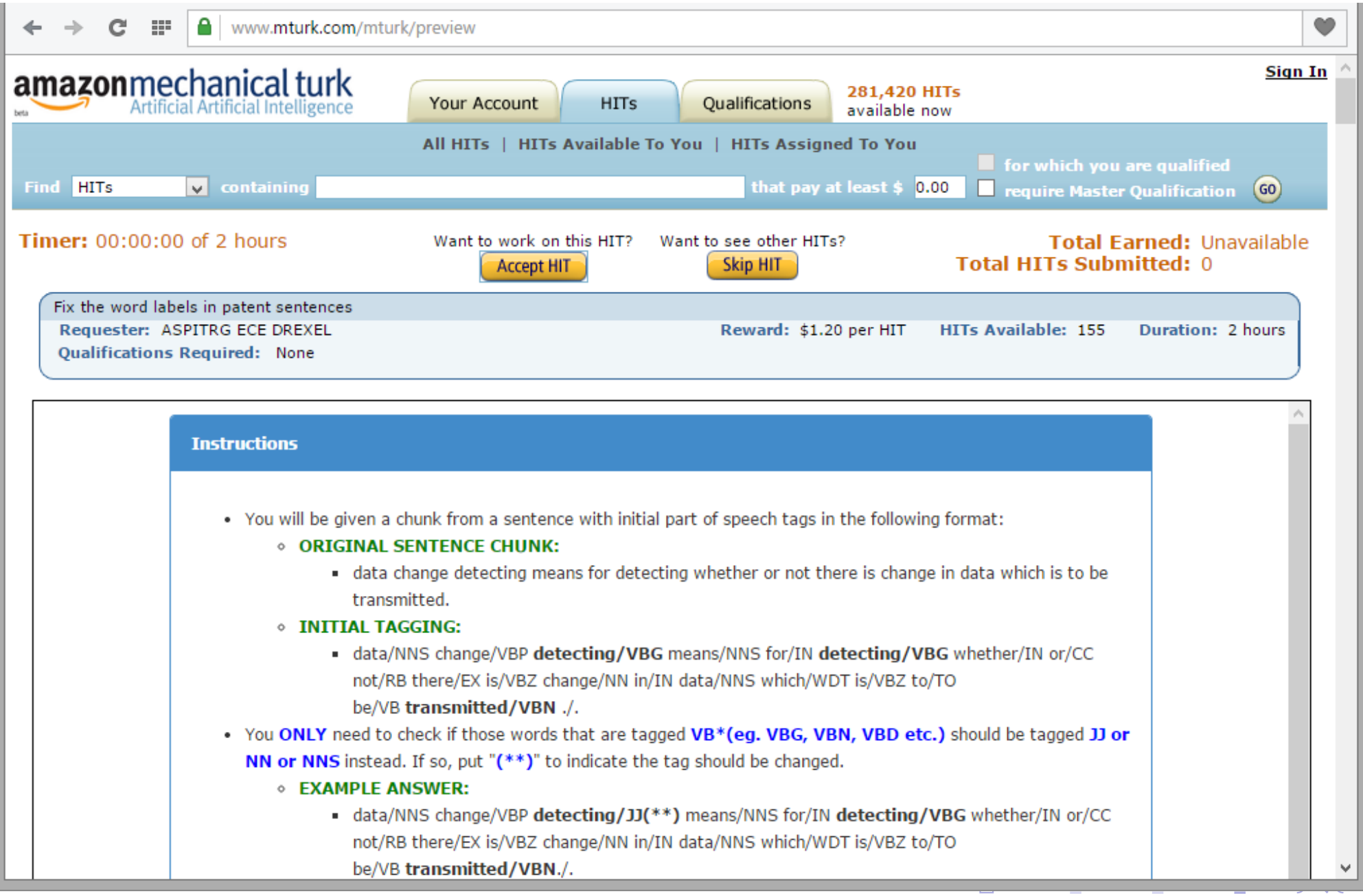

www.mturk.com/mturk/preview
amazonmechanical turk
Artificial Artificial Intelligence
Sign In
Your Account
HITs
Qualifications
281,420 HITs
available now
All HITs | HITs Available To You | HITs Assigned To You
Find HITs containing that pay at least $ 0.00
for which you are qualified
require Master Qualification
GO
Timer: 00:00:00 of 2 hours
Want to work on this HIT?
Accept HIT
Want to see other HITs?
Skip HIT
Total Earned: Unavailable
Total HITs Submitted: 0
Fix the word labels in patent sentences
Requester: ASPITRG ECE DREXEL
Reward: $1.20 per HIT
HITs Available: 155
Duration: 2 hours
Qualifications Required: None
Instructions
You will be given a chunk from a sentence with initial part of speech tags in the following format:
ORIGINAL SENTENCE CHUNK:
data change detecting means for detecting whether or not there is change in data which is to be transmitted.
INITIAL TAGGING:
data/NNS change/VBP detecting/VBG means/NNS for/IN detecting/VBG whether/IN or/CC not/RB there/EX is/VBZ change/NN in/IN data/NNS which/WDT is/VBZ to/TO be/VB transmitted/VBN ./.
You ONLY need to check if those words that are tagged VB*(eg. VBG, VBN, VBD etc.) should be tagged JJ or NN or NNS instead. If so, put "(**)" to indicate the tag should be changed.
EXAMPLE ANSWER:
data/NNS change/VBP detecting/JJ(**) means/NNS for/IN detecting/VBG whether/IN or/CC not/RB there/EX is/VBZ change/NN in/IN data/NNS which/WDT is/VBZ to/TO be/VB transmitted/VBN./.

Table of Contents

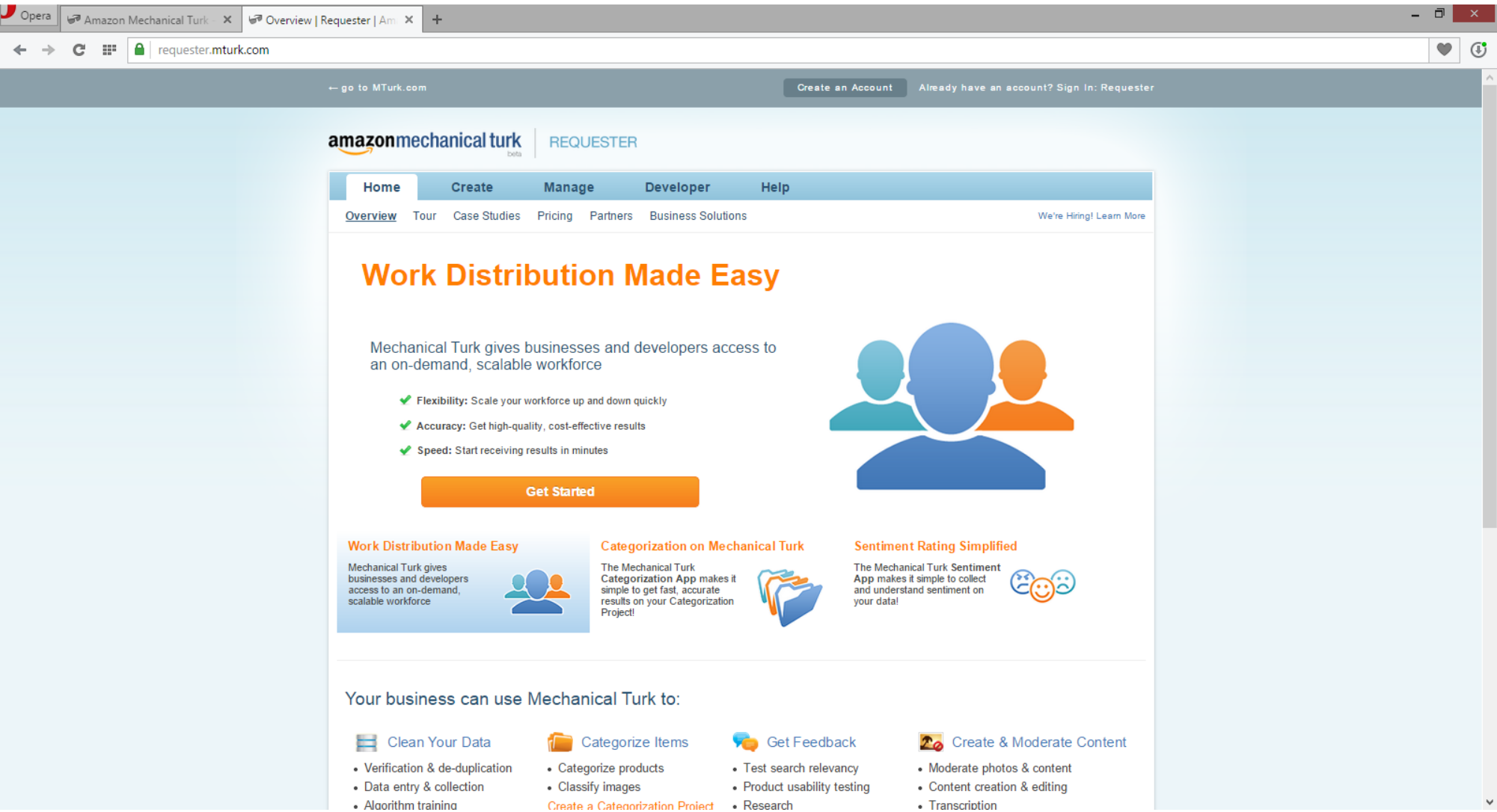
Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

Automatic Correcter of Patent Claim POS Tags

Conclusion

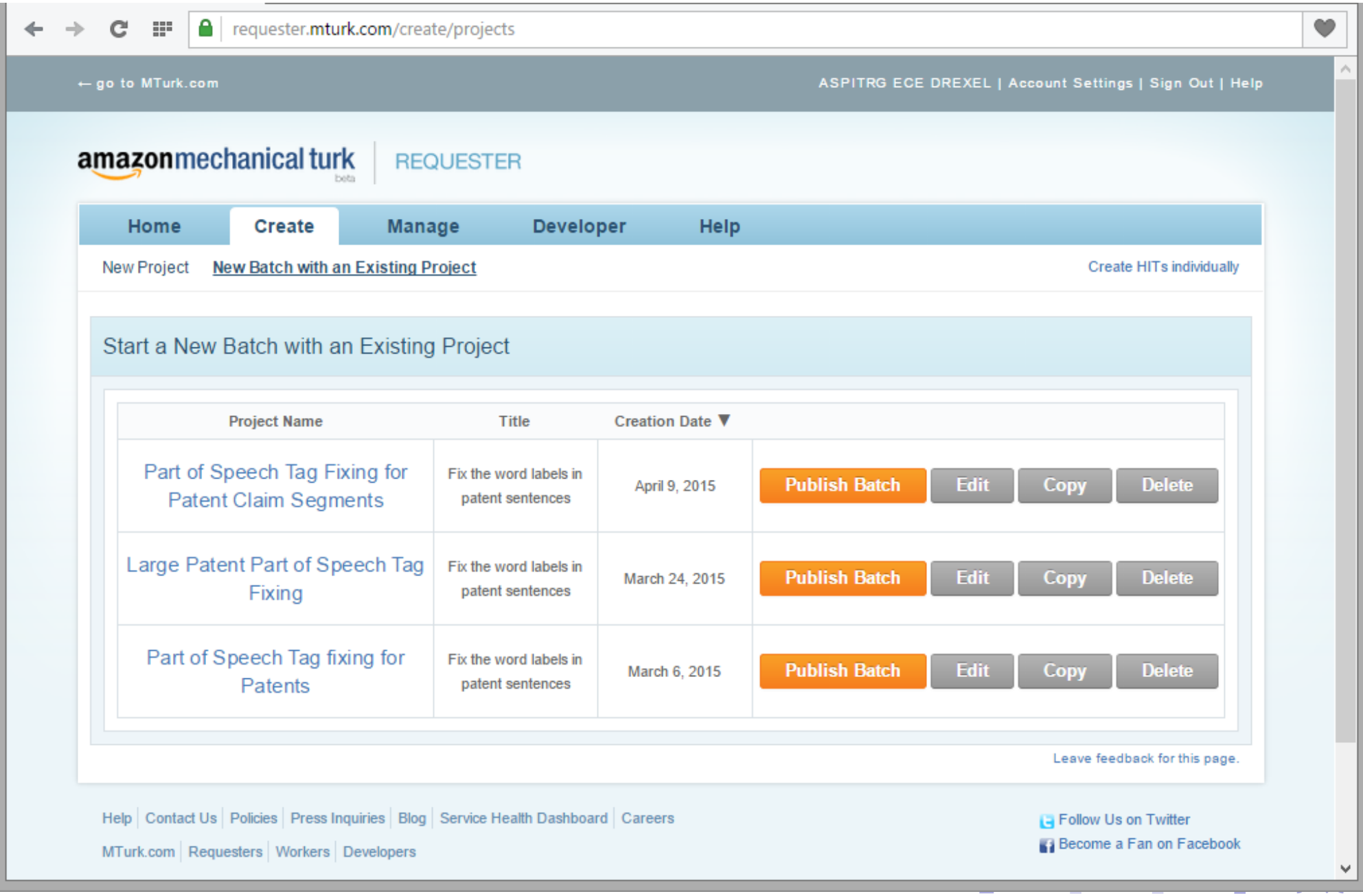
Requester Homepage



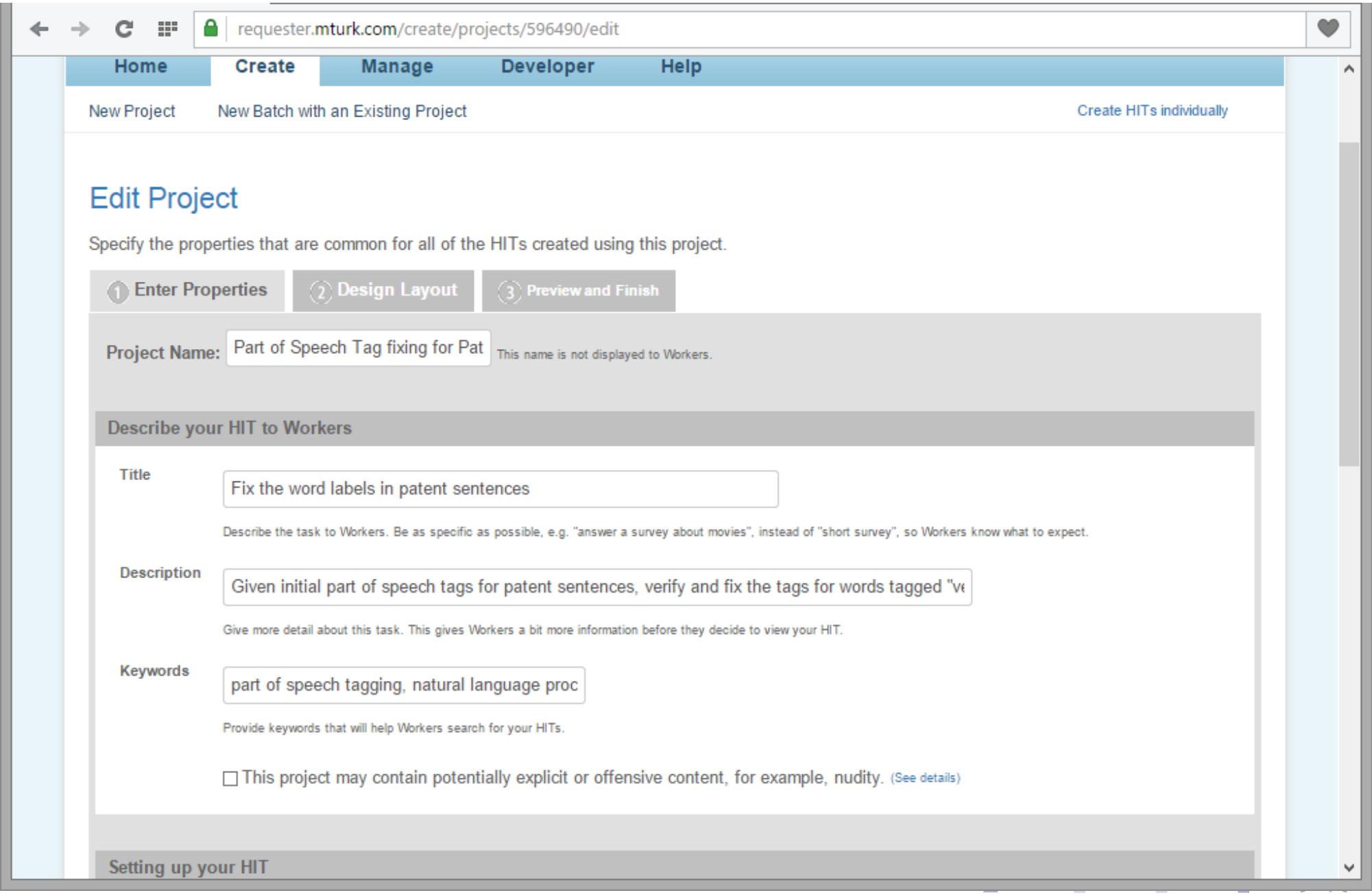
Tasks for Requesters

- **Editing HITs**
- Publishing Batches
- Reviewing Batches

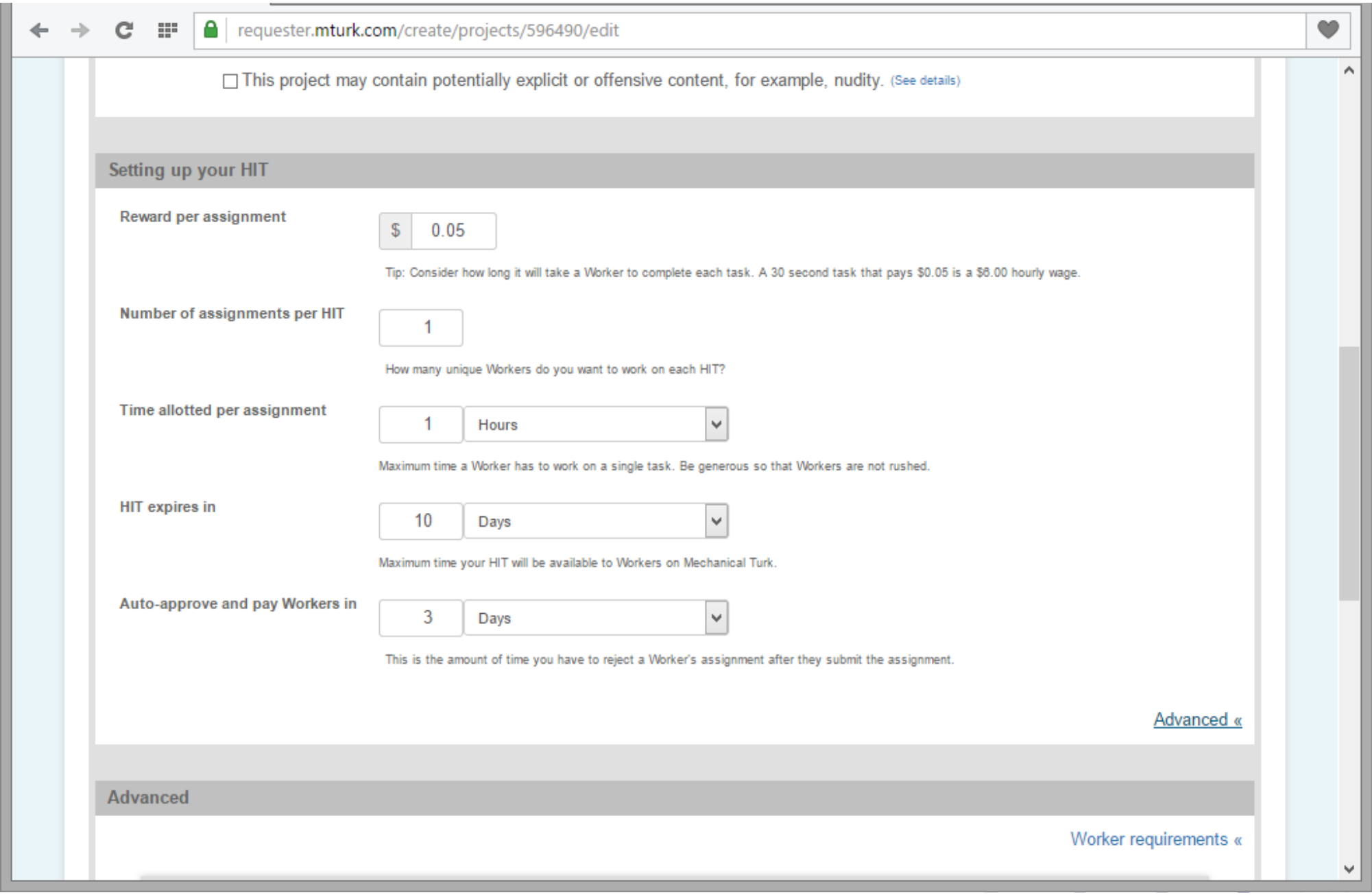
Editing HITs



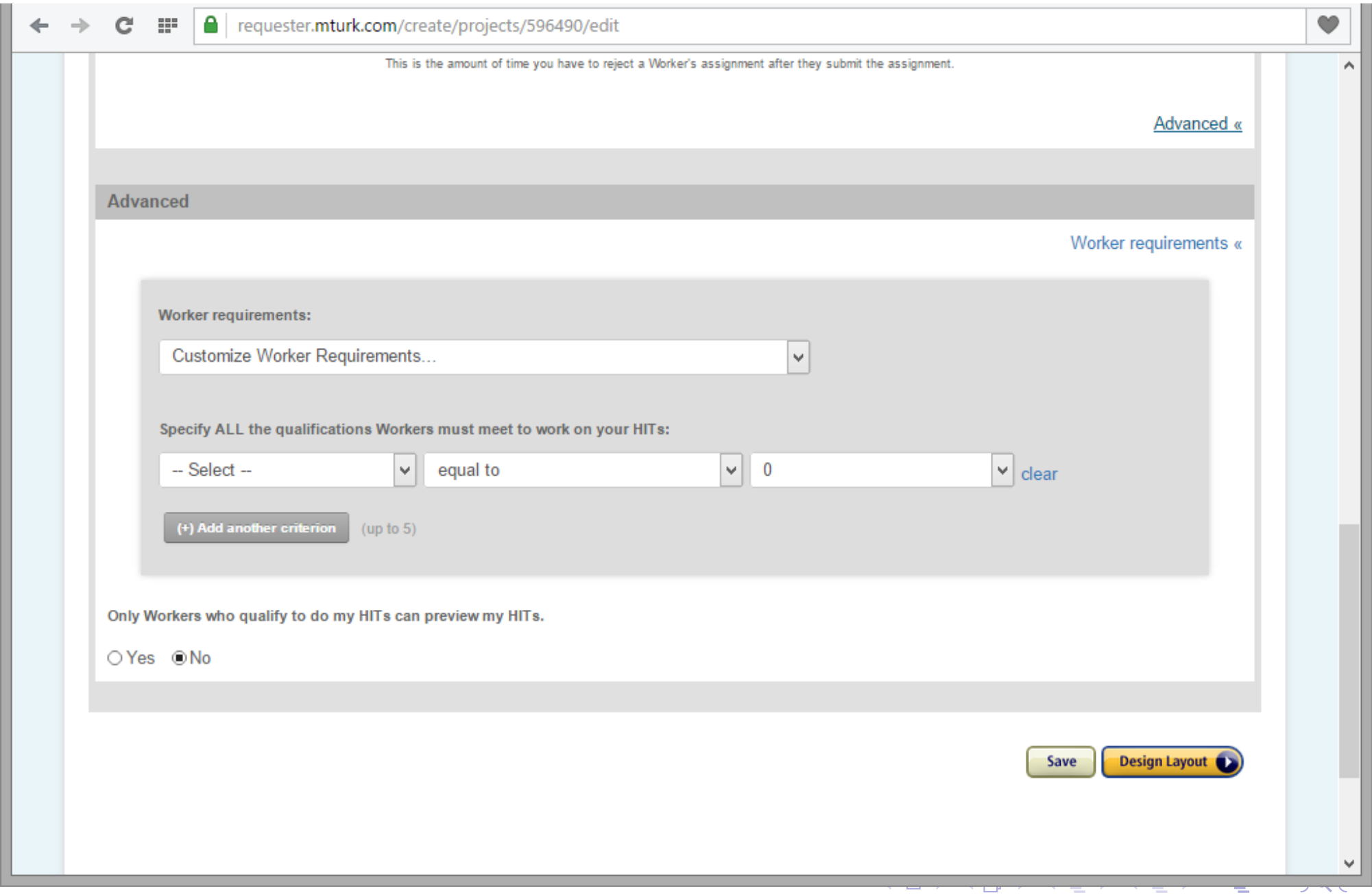
Editing HITs - Describing the HIT



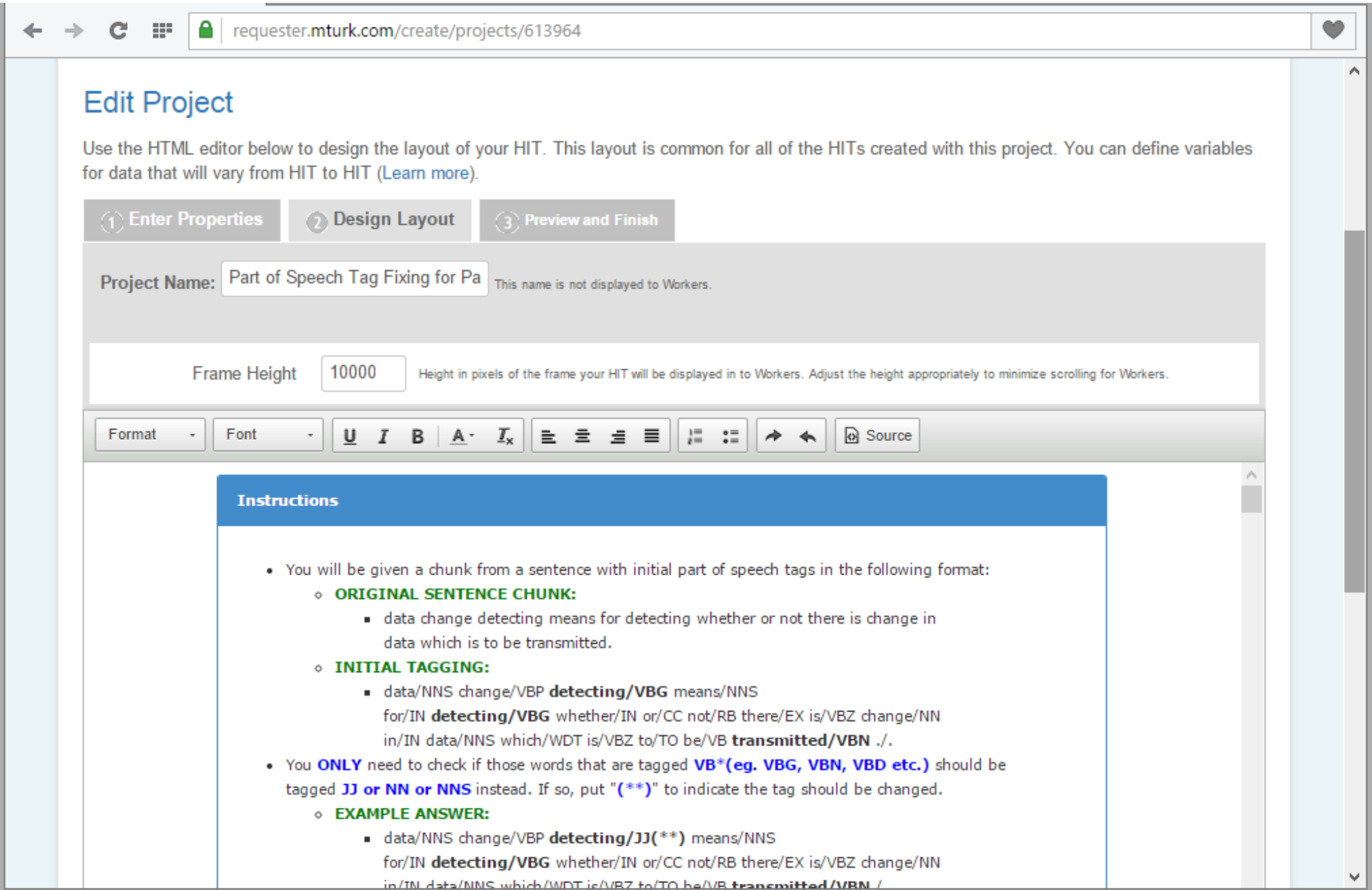
Editing HITs - Setting Up Properties



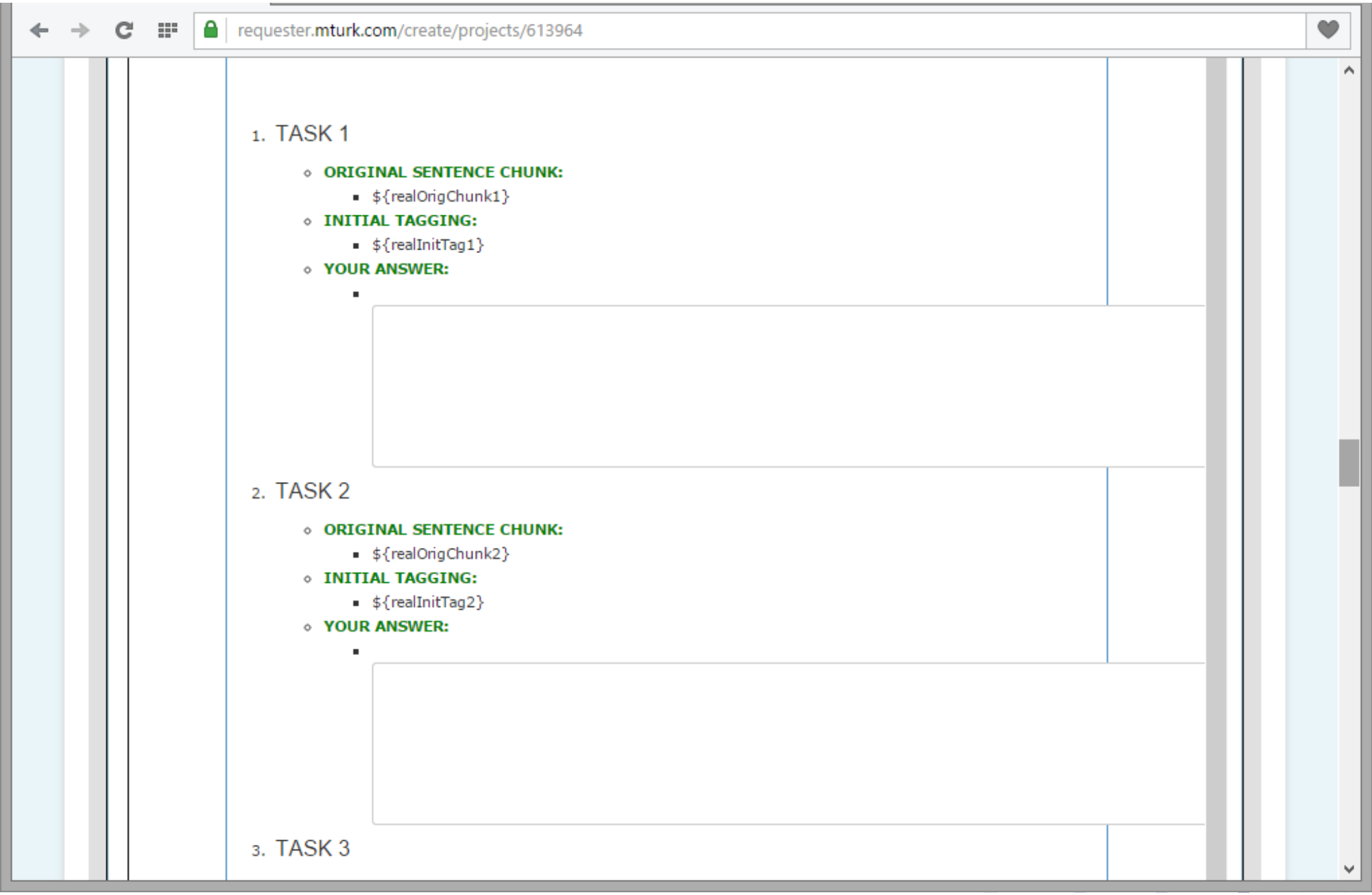
Editing HITs - Setting Worker Requirements



Editing HITs - Designing the HIT



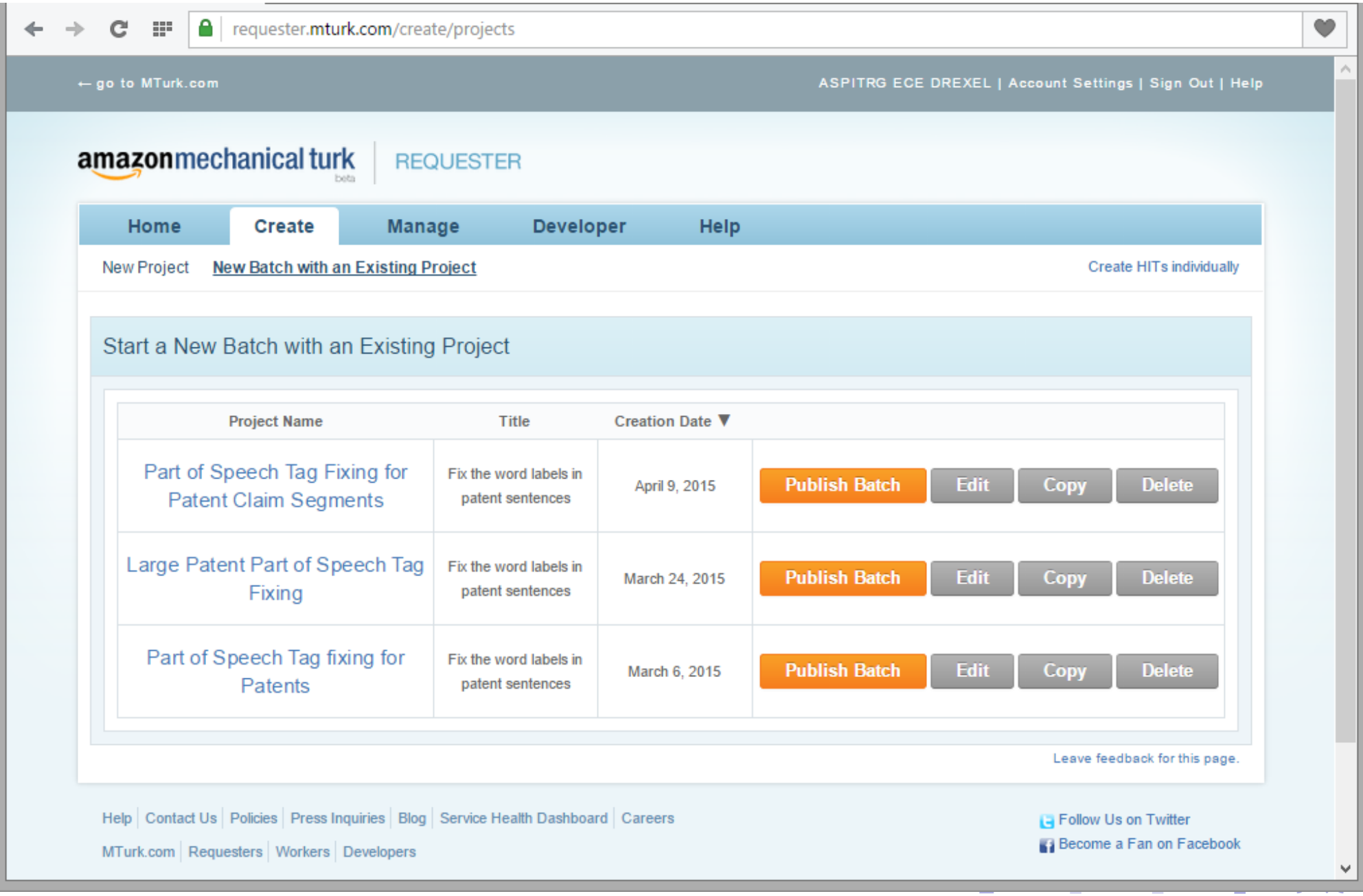
Editing HITs - Previewing the HIT



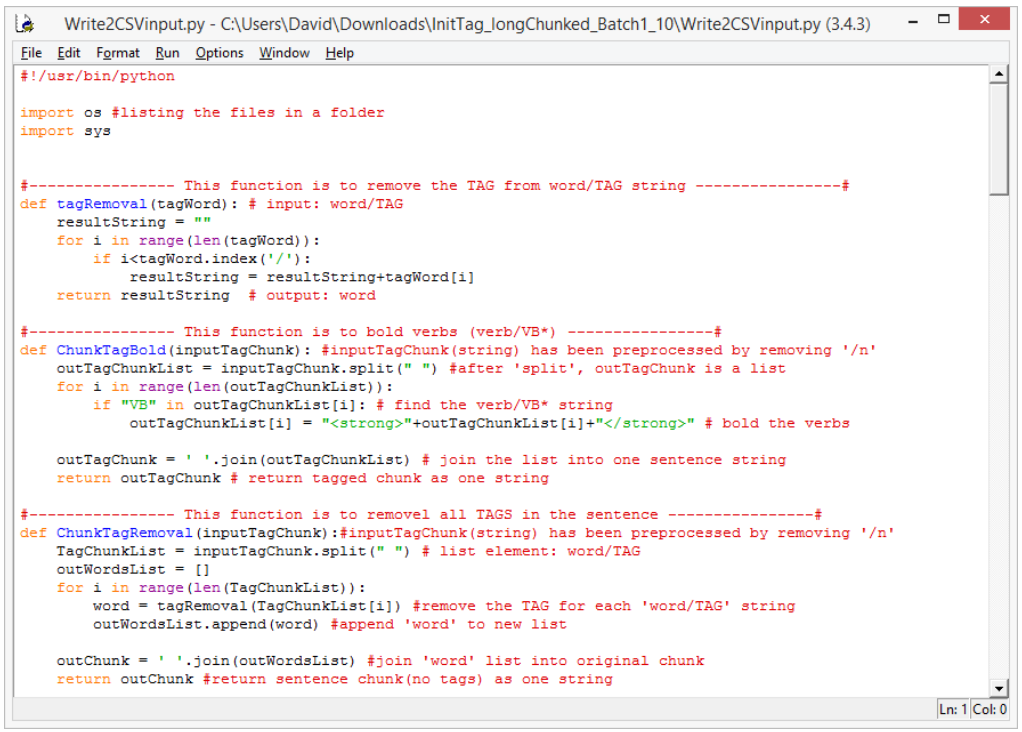
Tasks for Requesters

- Editing HITs
- **Publishing Batches**
- Reviewing Batches

Publishing HITs



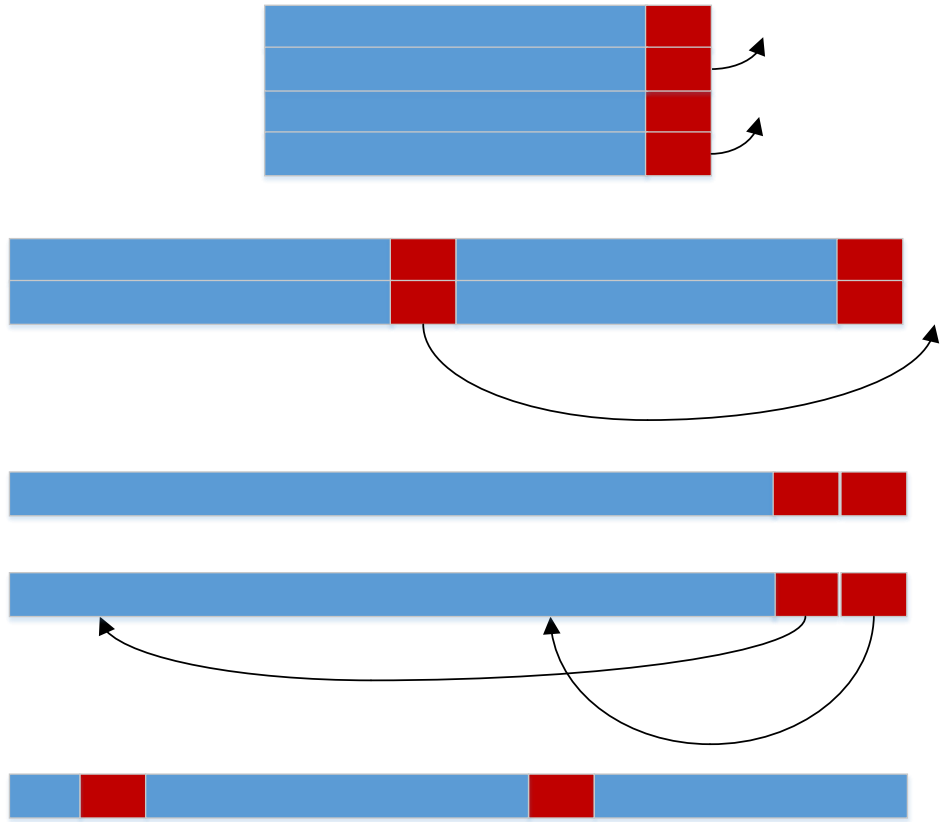
Publishing HITs - Developing the Initial CSV file



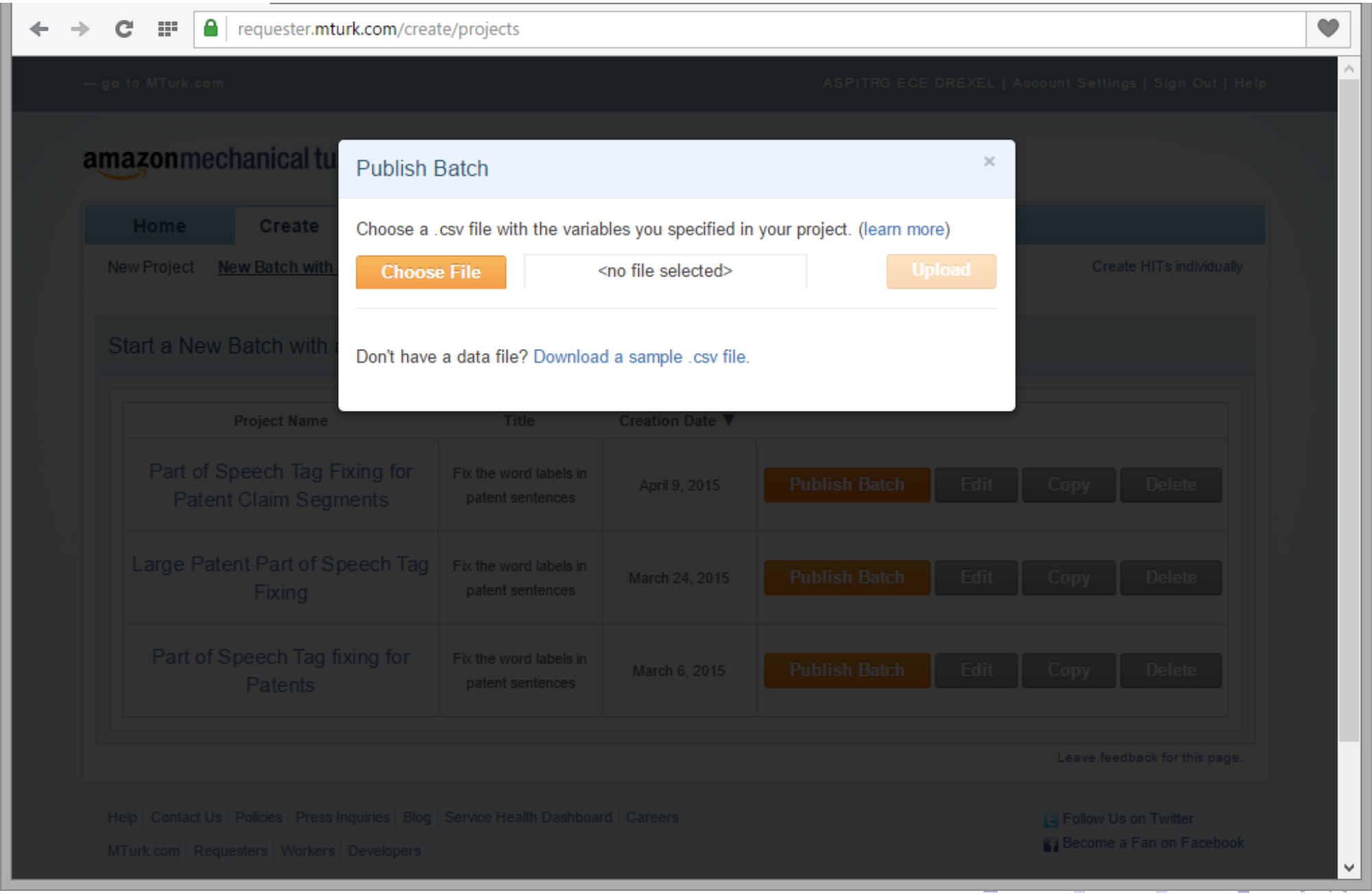
- Takes a folder with initial POS tagged patent claims that were not checked and a folder featuring correctly labeled claim segments
- Makes a CSV file featuring one segment (featuring verbs) from the initial patent claim files and one of the segments with an already correct segment
- Also outputs a file featuring all the correct segments
- From there we can extend it to a 10 segment version by hand

Publishing HITs - Developing a Larger CSV file

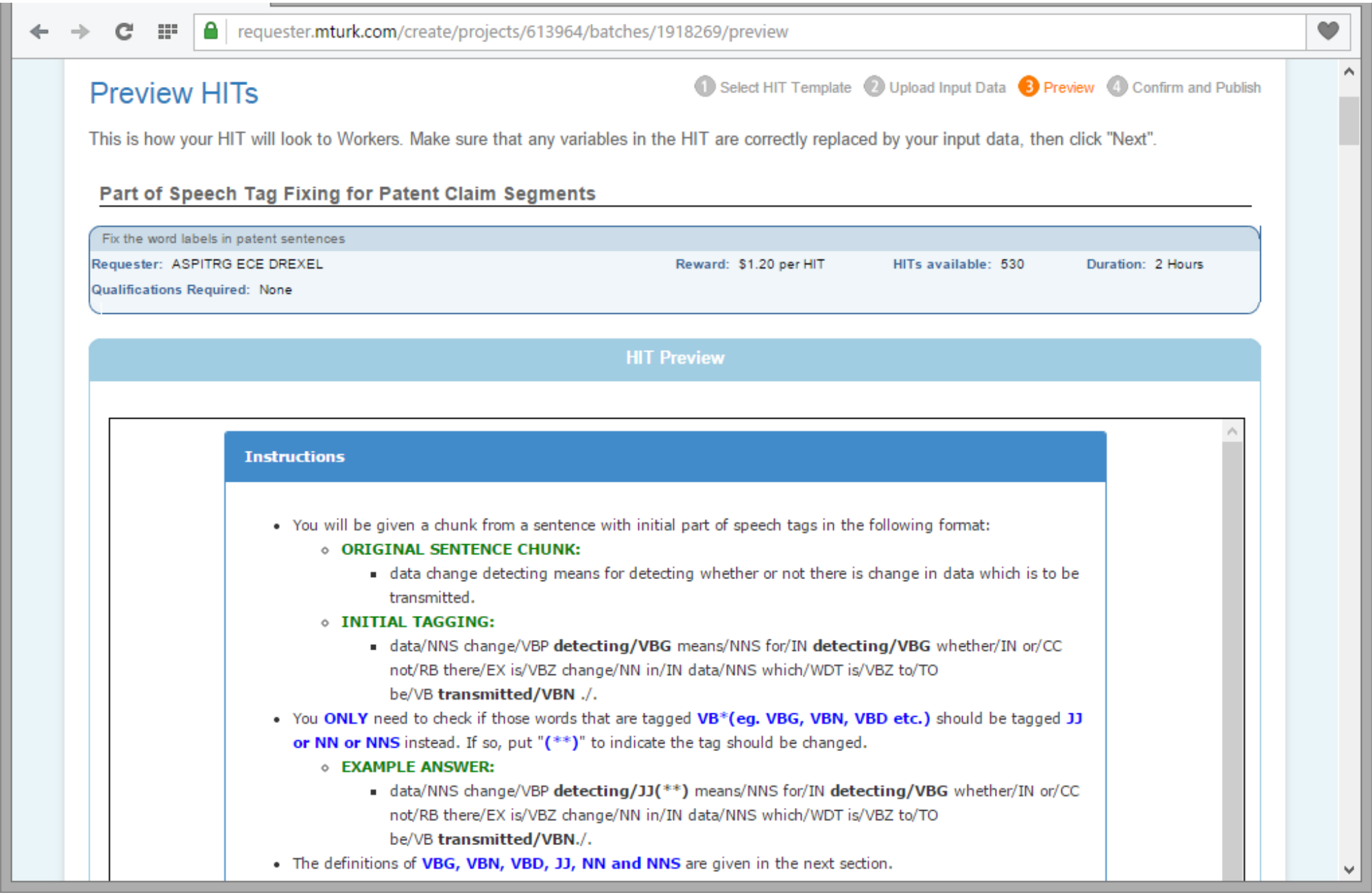
- VBA code that takes a 10 segment version and makes a 20 segment version
- Interleaves the segments of each HIT (including the test segments)
- Randomizes the locations of the two test segments and saves their locations as a hidden variable



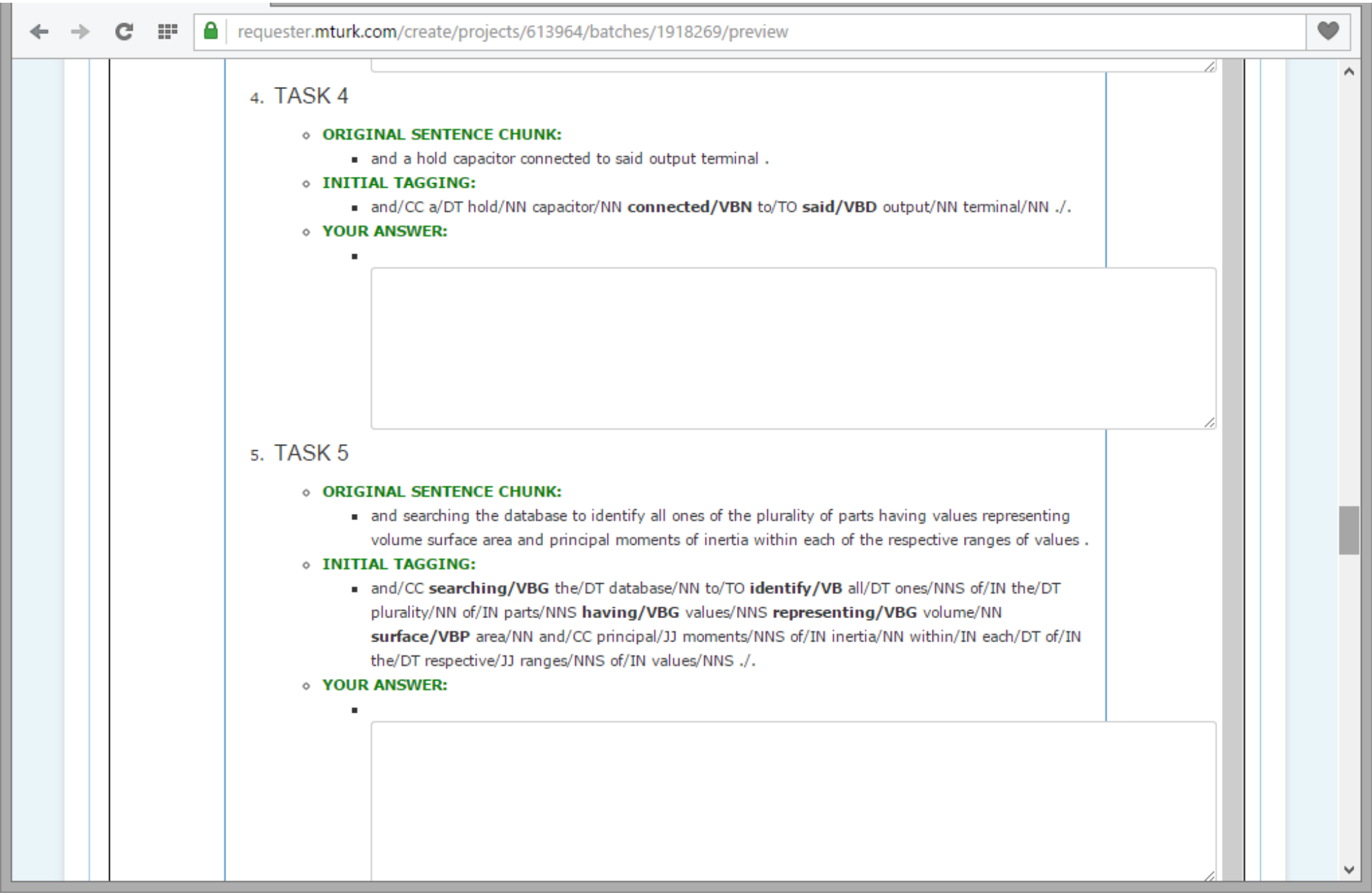
Publishing HITs - Uploading the CSV file



Publishing HITs - Previewing the HITs



Publishing HITs - Previewing the HITs



Publishing HITs - Confirming the Batch

requester.mturk.com/create/projects/613964/batches/1918269/checkout

Batch Summary

Batch Name: Part of Speech Tag Fixing for Pa

Description: Given initial part of speech tags f

Batch Properties

Title:	Fix the word labels in patent sentences
Description:	Given initial part of speech tags for patent sentences, verify and fix the tags for words tagged as "verb"
Batch expires in:	30 Days
Results are auto-approved and Workers are paid after:	5 Days

HITs

Number of HITs in this batch:		530
Number of assignments per HIT:	x	1
Total number of assignments in this batch:		530

Cost

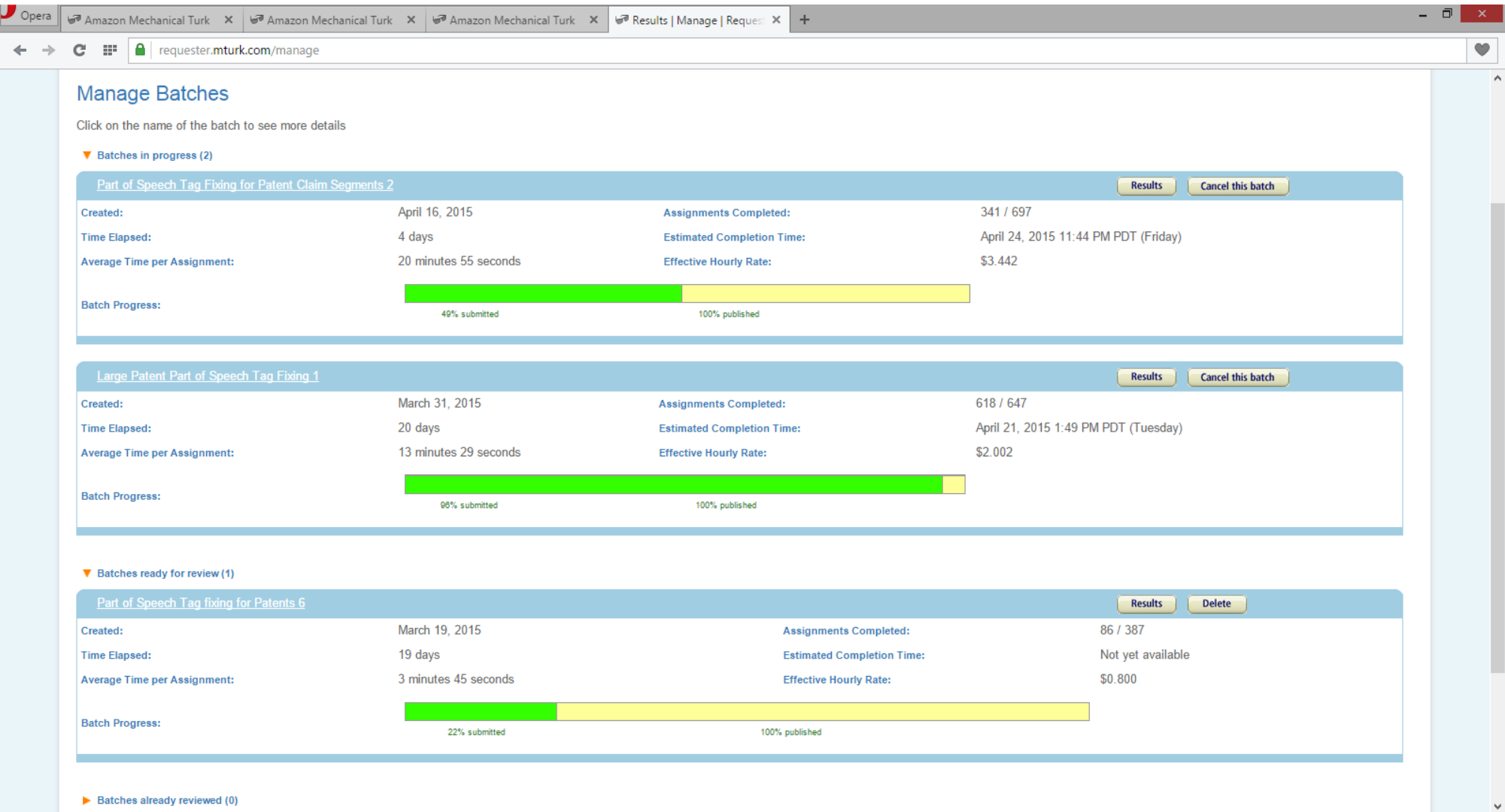
Reward per Assignment:		$1.200	
	x	530	*(total number of assignments in this batch)*
Estimated Total Reward:		$636.000	
Estimated Fees to Mechanical Turk:	+	$63.600	*(fees paid to Mechanical Turk) (fee details)*
Estimated Total Cost:		**$699.600**	*(this is the amount that will be deducted from your Available Balance when you click "Publish HITs")*

Your Available Balance:	$135.345	*(before clicking "Publish HITs")*
Your Projected Balance:	- $564.255	*(after clicking "Publish HITs")*

Tasks for Requesters

- Editing HITs
- Publishing Batches
- **Reviewing Batches**

Reviewing Batches



Reviewing Batches - Summary of Results

requester.mturk.com/batches/1904966

Given initial part of speech tags for patent sentences, verify and fix the tags for words tagged as "verb"

Status — Cancel this batch

Status: In Progress

83% submitted 100% published

Assignments Completed:	772 / 928	**Average Time per Assignment:**	19 minutes 28 seconds	**Average Hourly Rate:**	$3.699
Creation Time:	April 16, 2015 10:18 AM PDT	**Estimated Completion Time:**	April 28, 2015 4:57 AM PDT (Tuesday)		

Settings

Part of Speech Tag Fixing for Patent Claim Segments

Description:	Given initial part of speech tags for patent sentences, verify and fix the tags for words tagged as "verb"
Keywords:	part of speech tagging, natural language processing, patent, patent claims
Qualification Requirement:	
Number of Assignments per HIT:	1
Reward per Assignment:	$1.200
Input File:	largePatentTurkTask.csv
HIT expires on:	May 25, 2015 6:48 AM PDT (Monday)
Assignment duration:	2 hours
Auto Approval Delay:	5 days

Extend this batch

Results — Results

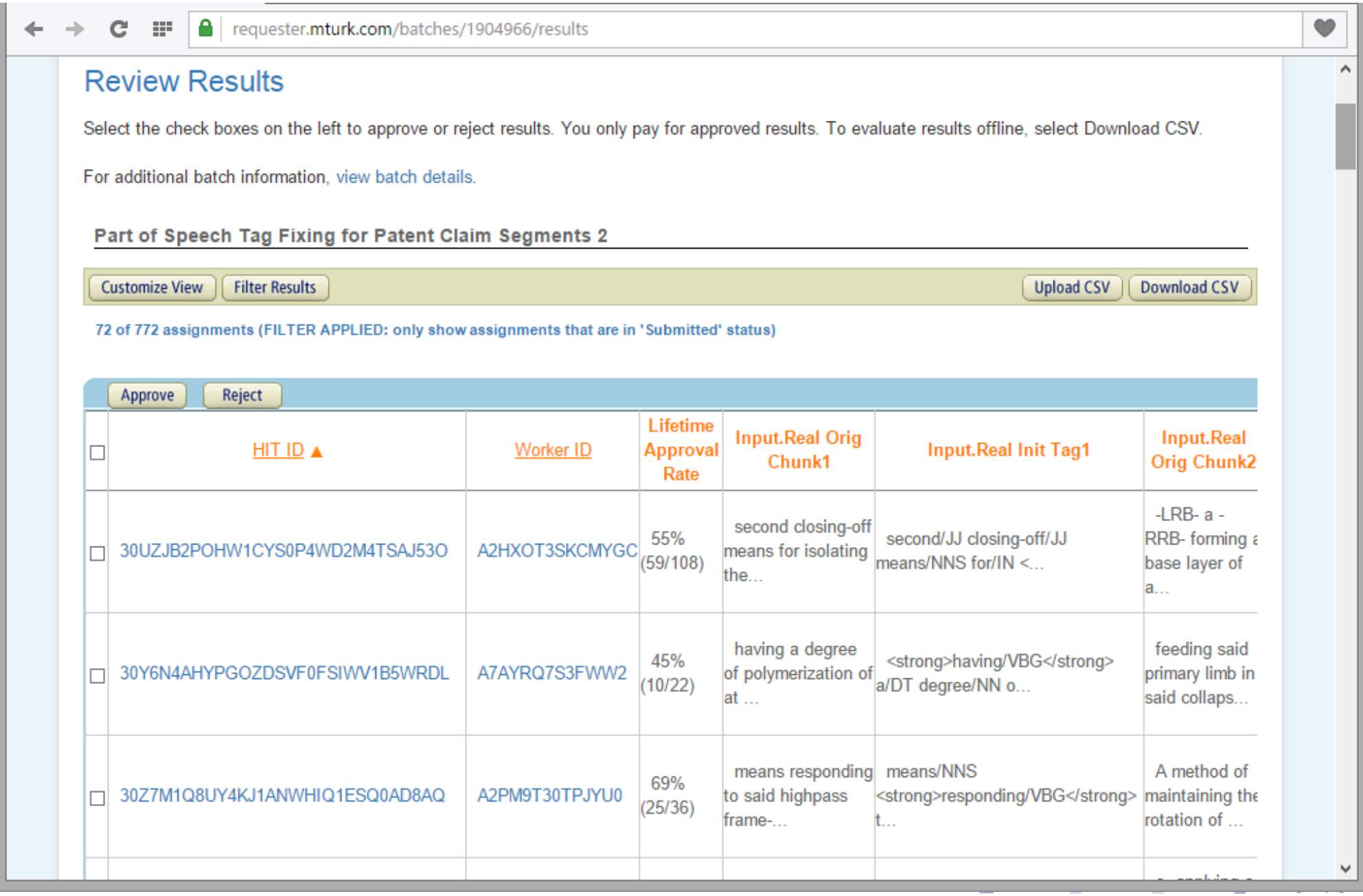
Assignments pending review:	72
Assignments approved:	308
Assignments rejected:	392

Cost Summary

Estimated Total Reward:	$636.000
Estimated Fees to Mechanical Turk:	$63.600 (fee details)
Estimated Total Cost:	$699.600

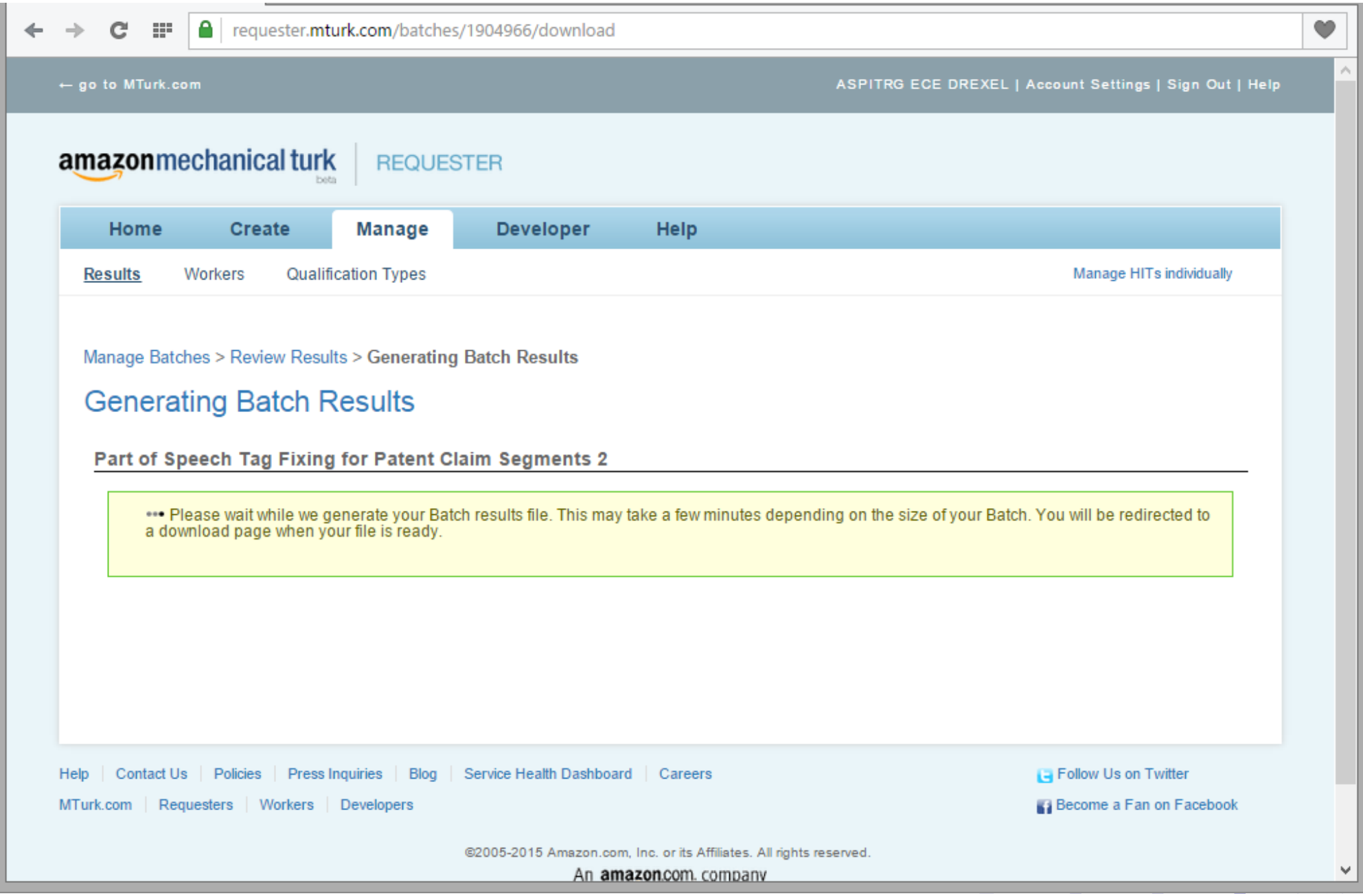
These costs are only an estimate until all of the assignments have been submitted and reviewed.

Reviewing Batches - Reviewing Results

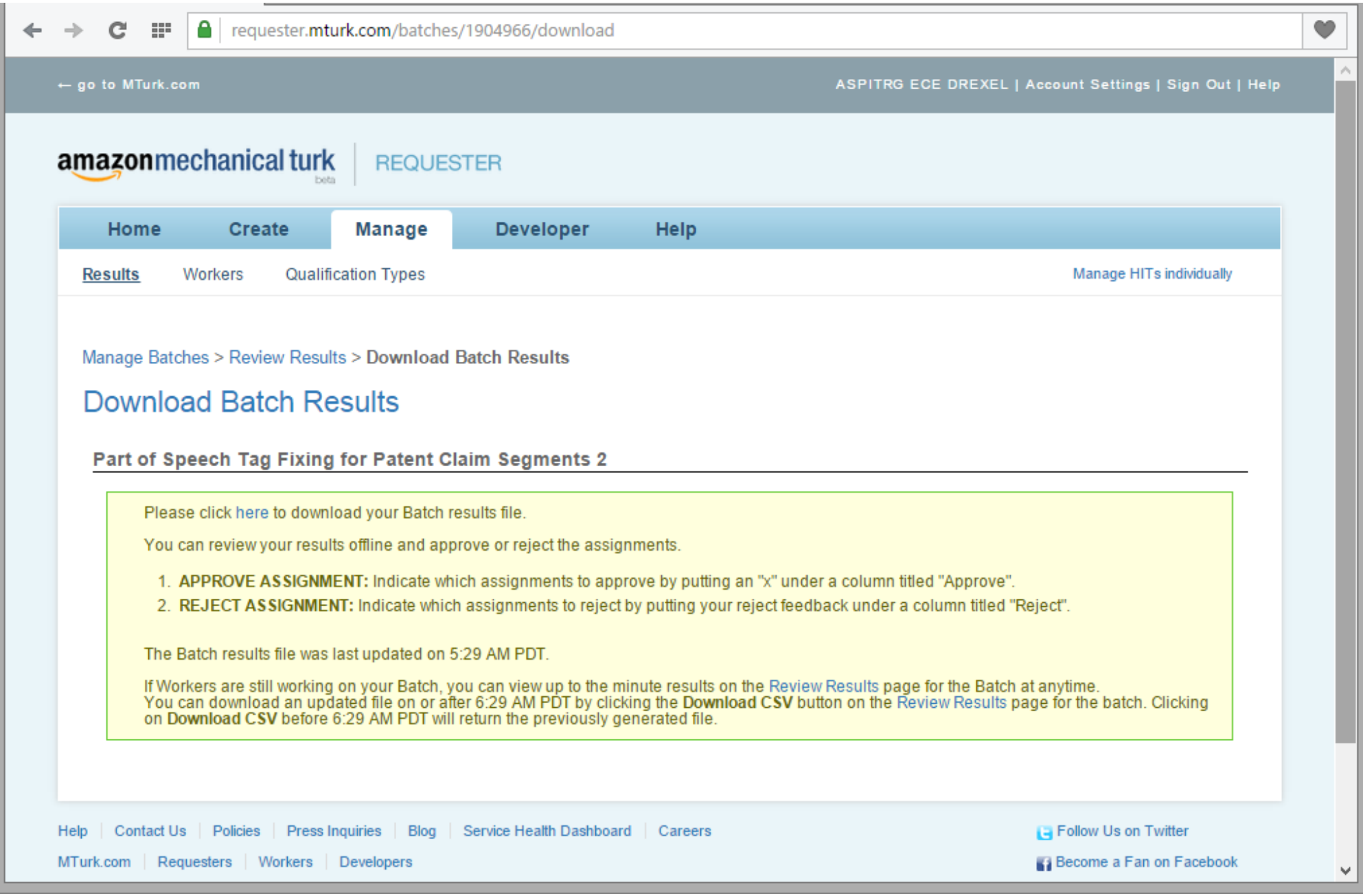


HIT ID ▲	Worker ID	Lifetime Approval Rate	Input.Real Orig Chunk1	Input.Real Init Tag1	Input.Real Orig Chunk2
30UZJB2POHW1CYS0P4WD2M4TSAJ53O	A2HXOT3SKCMYGC	55% (59/108)	second closing-off means for isolating the...	second/JJ closing-off/JJ means/NNS for/IN <...	-LRB- a -RRB- forming a base layer of a...
30Y6N4AHYPGOZDSVF0FSIWV1B5WRDL	A7AYRQ7S3FWW2	45% (10/22)	having a degree of polymerization of at ...	having/VBG a/DT degree/NN o...	feeding said primary limb in said collaps...
30Z7M1Q8UY4KJ1ANWHIQ1ESQ0AD8AQ	A2PM9T30TPJYU0	69% (25/36)	means responding to said highpass frame-...	means/NNS responding/VBG t...	A method of maintaining the rotation of ...

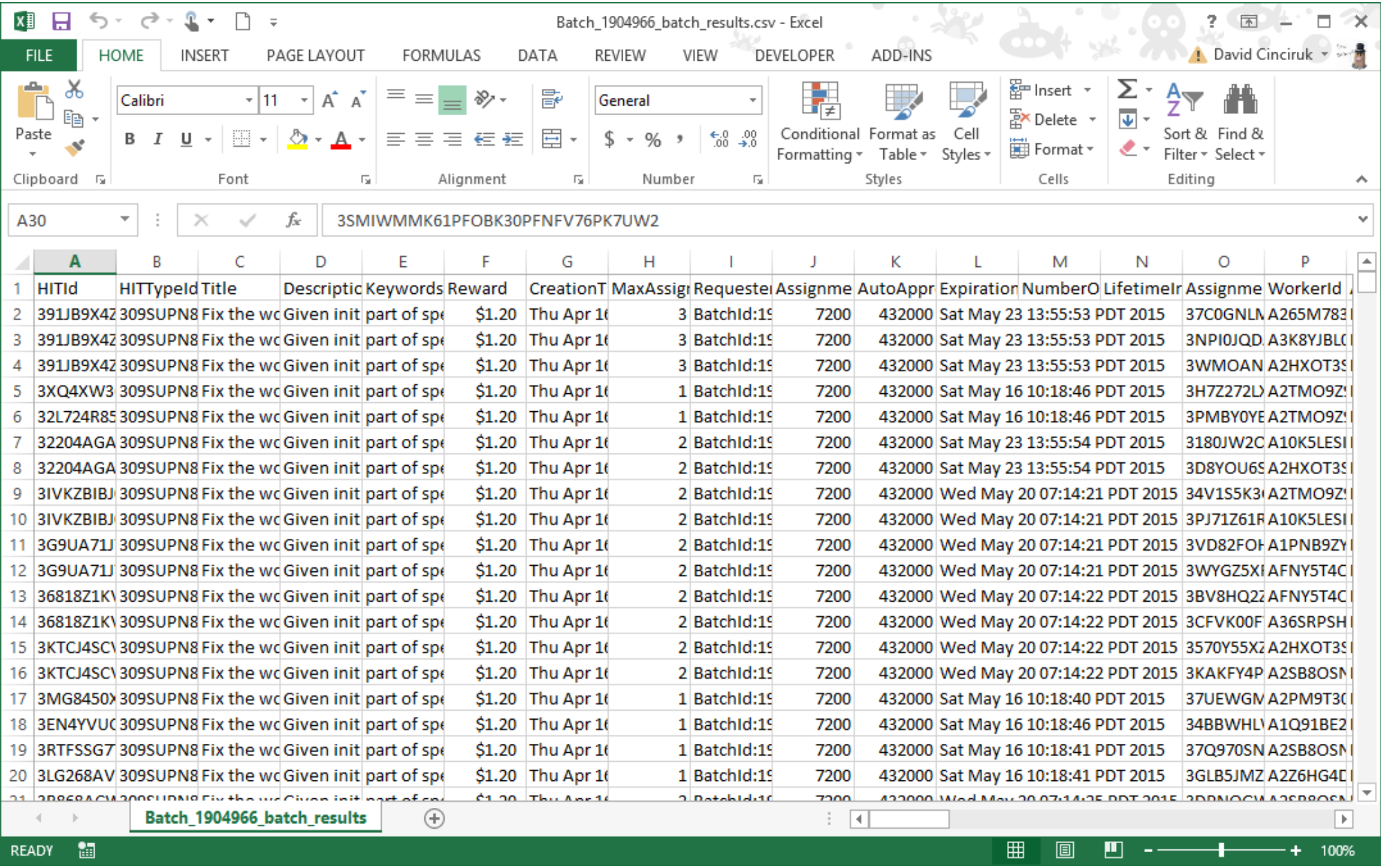
Reviewing Batches - Reviewing Results



Reviewing Batches - Reviewing Results

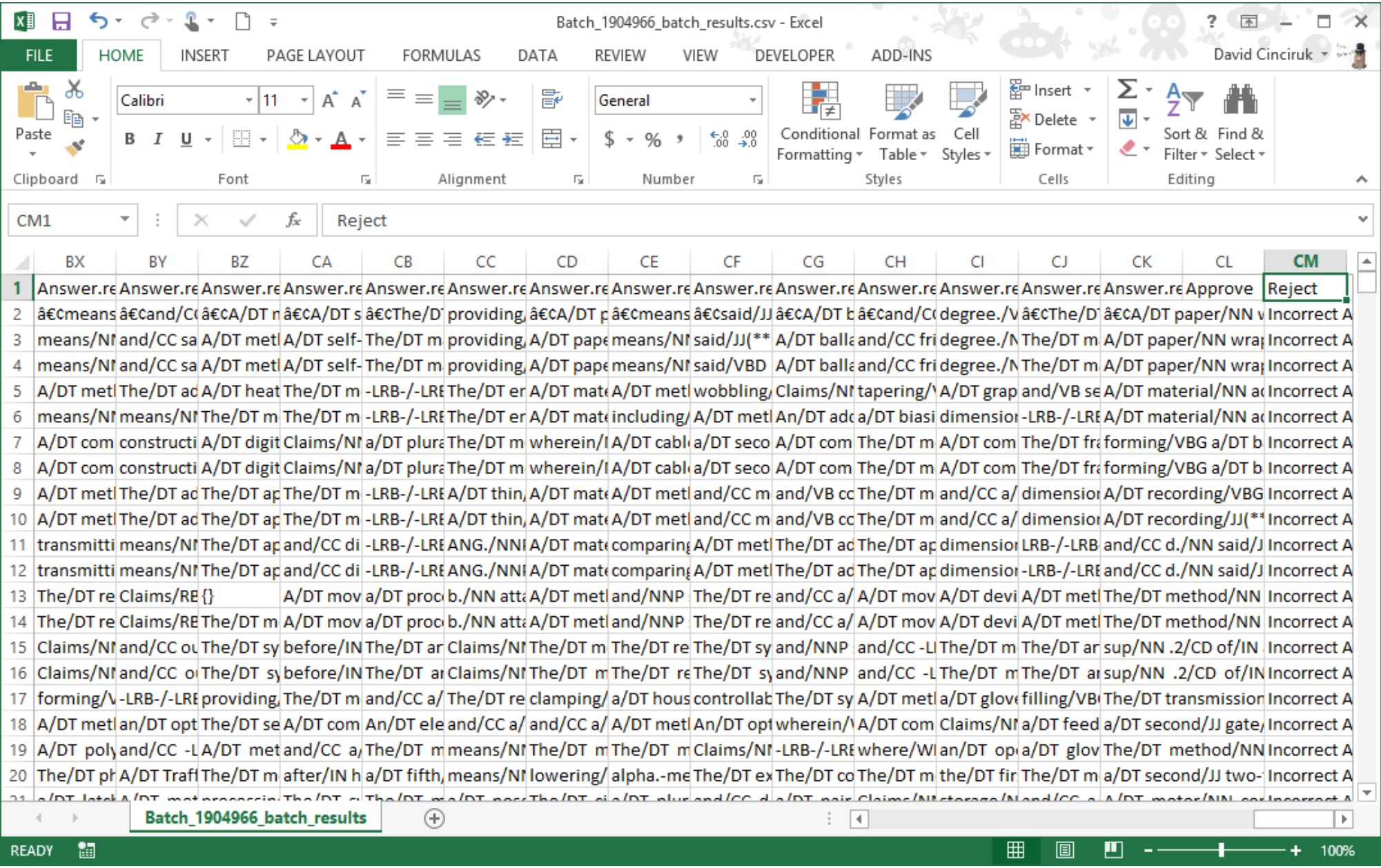


Reviewing Batches - CSV File



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	HITId	HITTypeId	Title	Descriptic	Keywords	Reward	CreationT	MaxAssigr	Requeste	Assignme	AutoAppr	Expiration	NumberO	LifetimeIr	Assignme	WorkerId
2	391JB9X4Z	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	3	BatchId:19	7200	432000	Sat May 23 13:55:53 PDT 2015			37C0GNLN	A265M783
3	391JB9X4Z	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	3	BatchId:19	7200	432000	Sat May 23 13:55:53 PDT 2015			3NPI0JQD	A3K8YJBL0
4	391JB9X4Z	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	3	BatchId:19	7200	432000	Sat May 23 13:55:53 PDT 2015			3WMOAN	A2HXOT3S
5	3XQ4XW3	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	1	BatchId:19	7200	432000	Sat May 16 10:18:46 PDT 2015			3H7Z272L	A2TMO9Z5
6	32L724R85	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	1	BatchId:19	7200	432000	Sat May 16 10:18:46 PDT 2015			3PMBY0YE	A2TMO9Z5
7	32204AGA	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Sat May 23 13:55:54 PDT 2015			3180JW2C	A10K5LESI
8	32204AGA	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Sat May 23 13:55:54 PDT 2015			3D8YOU6S	A2HXOT3S
9	3IVKZBIBJ	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:21 PDT 2015			34V1S5K3	A2TMO9Z5
10	3IVKZBIBJ	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:21 PDT 2015			3PJ71Z61R	A10K5LESI
11	3G9UA71J	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:21 PDT 2015			3VD82FOH	A1PNB9ZY
12	3G9UA71J	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:21 PDT 2015			3WYGZ5XI	AFNY5T4C
13	36818Z1K	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:22 PDT 2015			3BV8HQ2Z	AFNY5T4C
14	36818Z1K	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:22 PDT 2015			3CFVK00F	A36SRPSH
15	3KTCJ4SC	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:22 PDT 2015			3570Y55X	A2HXOT3S
16	3KTCJ4SC	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	2	BatchId:19	7200	432000	Wed May 20 07:14:22 PDT 2015			3KAKFY4P	A2SB8OSN
17	3MG8450X	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	1	BatchId:19	7200	432000	Sat May 16 10:18:40 PDT 2015			37UEWGM	A2PM9T3C
18	3EN4YVUC	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	1	BatchId:19	7200	432000	Sat May 16 10:18:46 PDT 2015			34BBWHLV	A1Q91BE2
19	3RTFSSG7	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	1	BatchId:19	7200	432000	Sat May 16 10:18:41 PDT 2015			37Q970SN	A2SB8OSN
20	3LG268AV	309SUPN8	Fix the wc	Given init	part of spe	$1.20	Thu Apr 16	1	BatchId:19	7200	432000	Sat May 16 10:18:41 PDT 2015			3GLB5JMZ	A2Z6HG4D

Reviewing Batches - CSV File



Reviewing Batches - Checking Answers

- Need to approve or reject HITs since answers may or may not be correct.
- May opt to give the same HIT to multiple Workers and then pay the majority answer
  - Better for subjective HITs
  - More Expensive since need to pay for multiple "correct" HITs
- May instead put test questions inside HITs that have known answers and check the answers
  - Better for objective HITs where accuracy is key
  - Method we used for our HITs

Reviewing Batches - Automatic Matlab HIT Checker

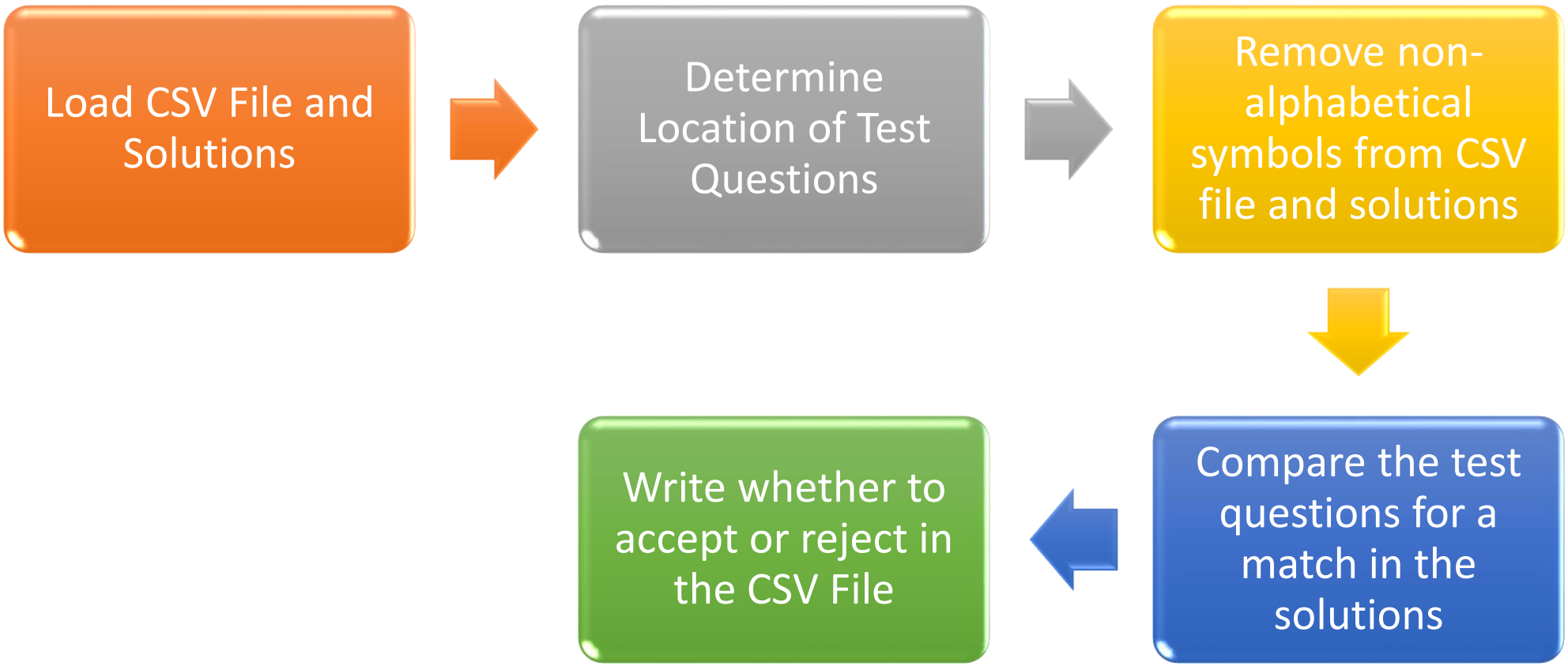


Table of Contents

Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

Automatic Correcter of Patent Claim POS Tags

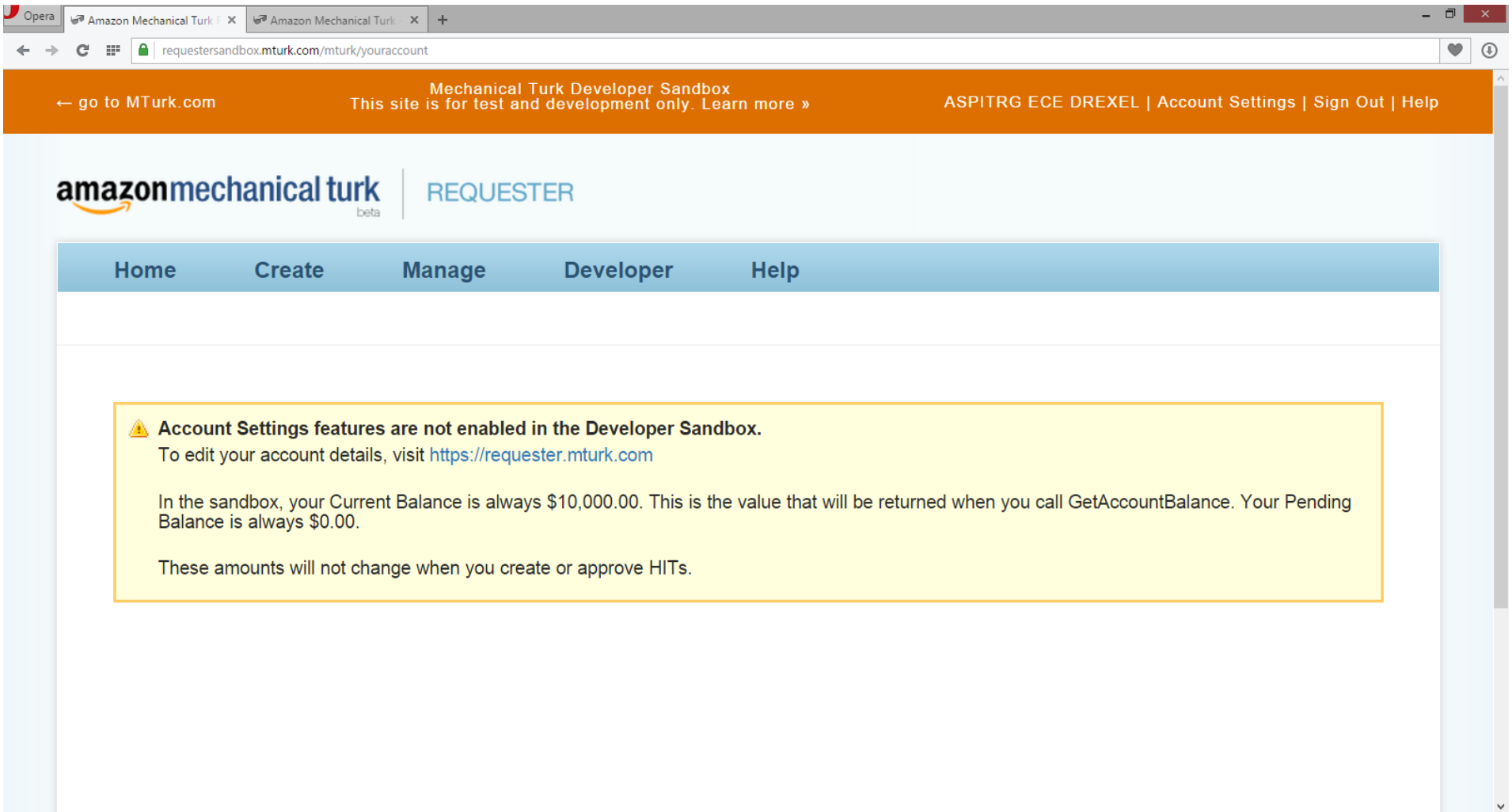
Conclusion

Mechanical Turk Sandbox

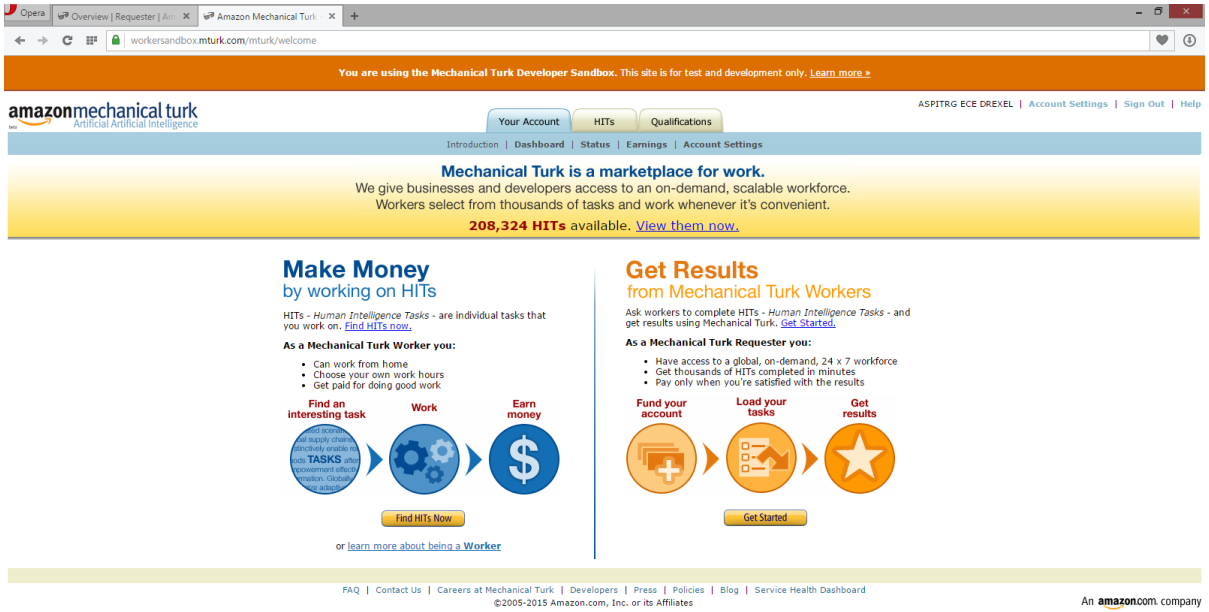
- You may not know how to develop your HIT or how your results will look.
- Sandbox mode allows Requesters to develop HITs without having to pay anyone.
- Requesters can make Sandbox Worker accounts to perform their own HITs to test out answering the HITs.
- Layout for Workers and Requesters are same as regular Turk

Like Monopoly Money

You have fake money on the Sandbox that you can give out to Sandbox Workers



Worker and Requester Sandbox



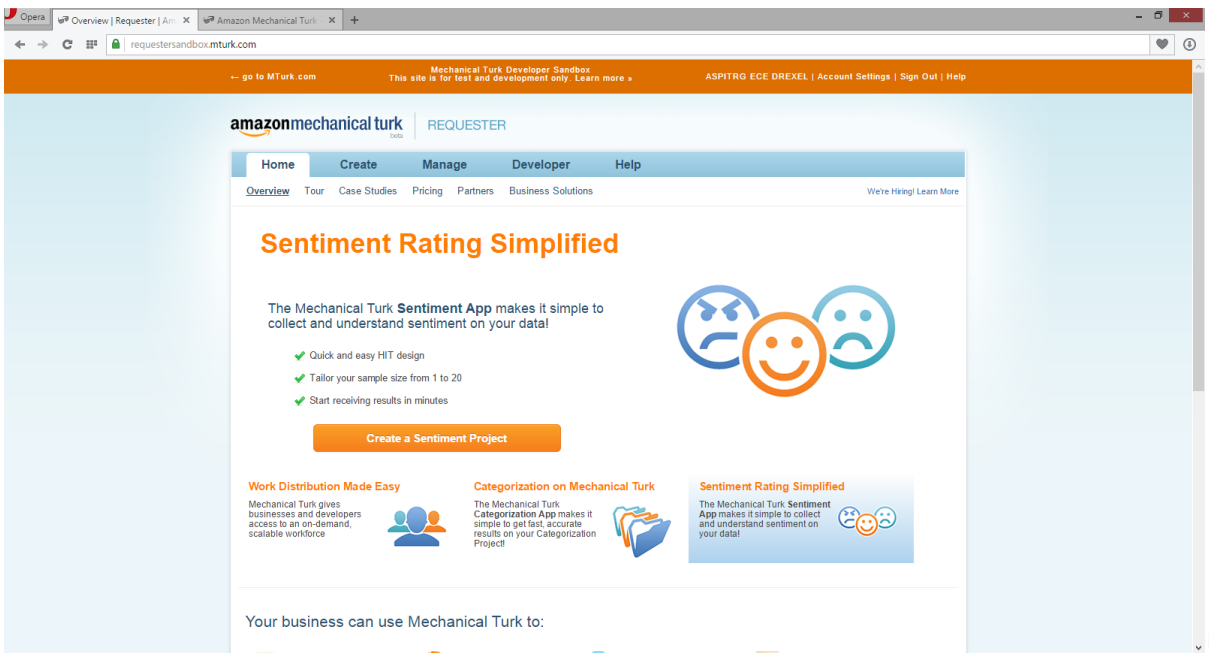


Table of Contents

Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

Automatic Correcter of Patent Claim POS Tags

Conclusion

Fishing for Results

Workers need to be enticed to do HITs. Many potential ways to do that. Need to weigh one lure against another

Our Experiences

Part of Speech Tag fixing for Patents	Fix the word labels in patent sentences	March 6, 2015	Publish Batch Edit Copy Delete

Initial unsucessful HIT paid 5 cents for 2 questions - after 19 days only 86 HITs completed

Large Patent Part of Speech Tag Fixing	Fix the word labels in patent sentences	March 24, 2015	Publish Batch Edit Copy Delete

Secondary successful HIT paid 45 cents for 10 questions - after 27 days 657 HITs completed

Part of Speech Tag Fixing for Patent Claim Segments	Fix the word labels in patent sentences	April 9, 2015	Publish Batch Edit Copy Delete

Very successful third HIT paid $1.20 for 20 questions - after 11 days 878 HITs completed

Our Experiences

- Higher Paying HITs entice more people
- More available HITs entice more people
- Reducing Qualifications allow more people to participate

Table of Contents

Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

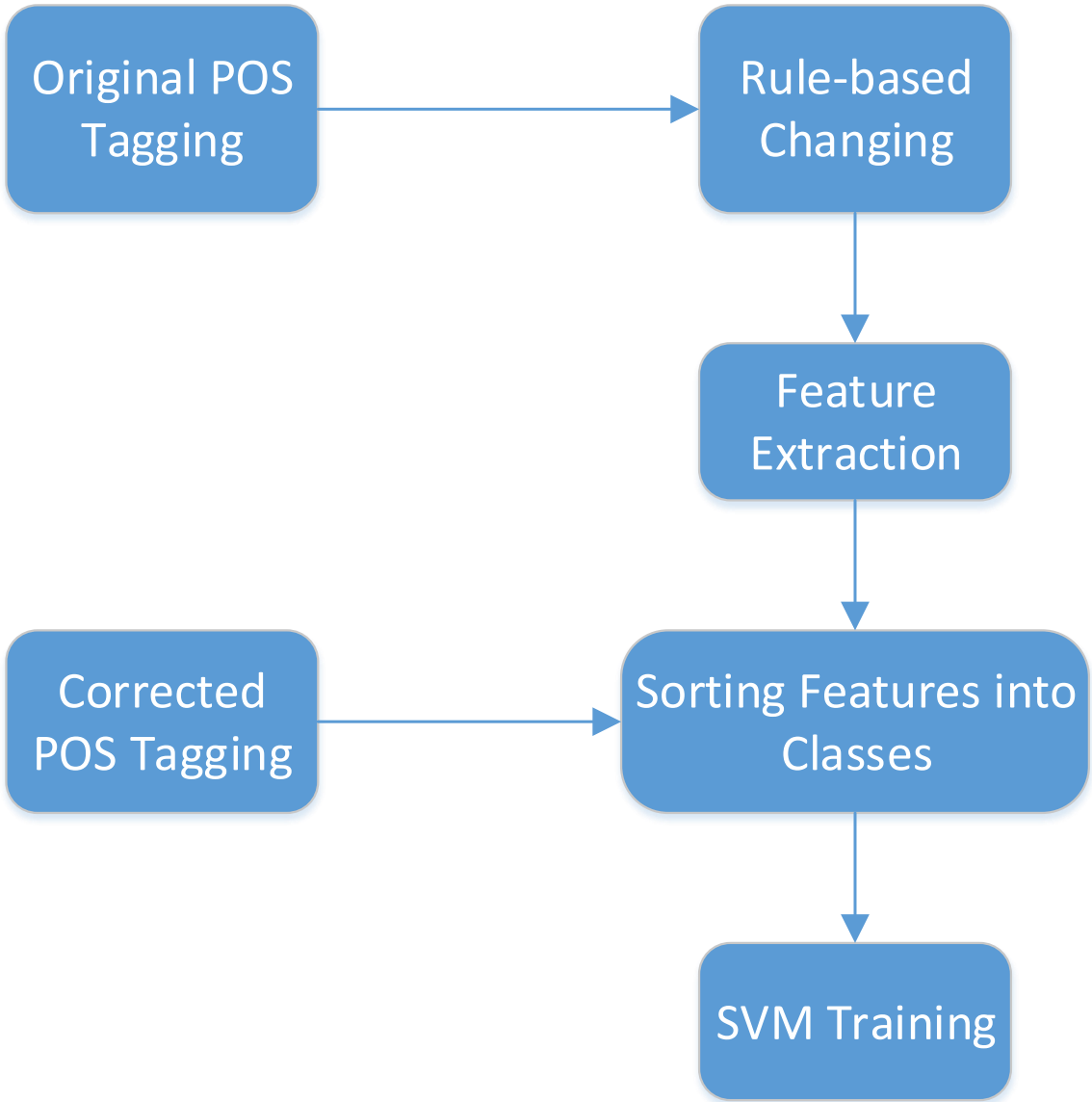
Automatic Correcter of Patent Claim POS Tags

Conclusion

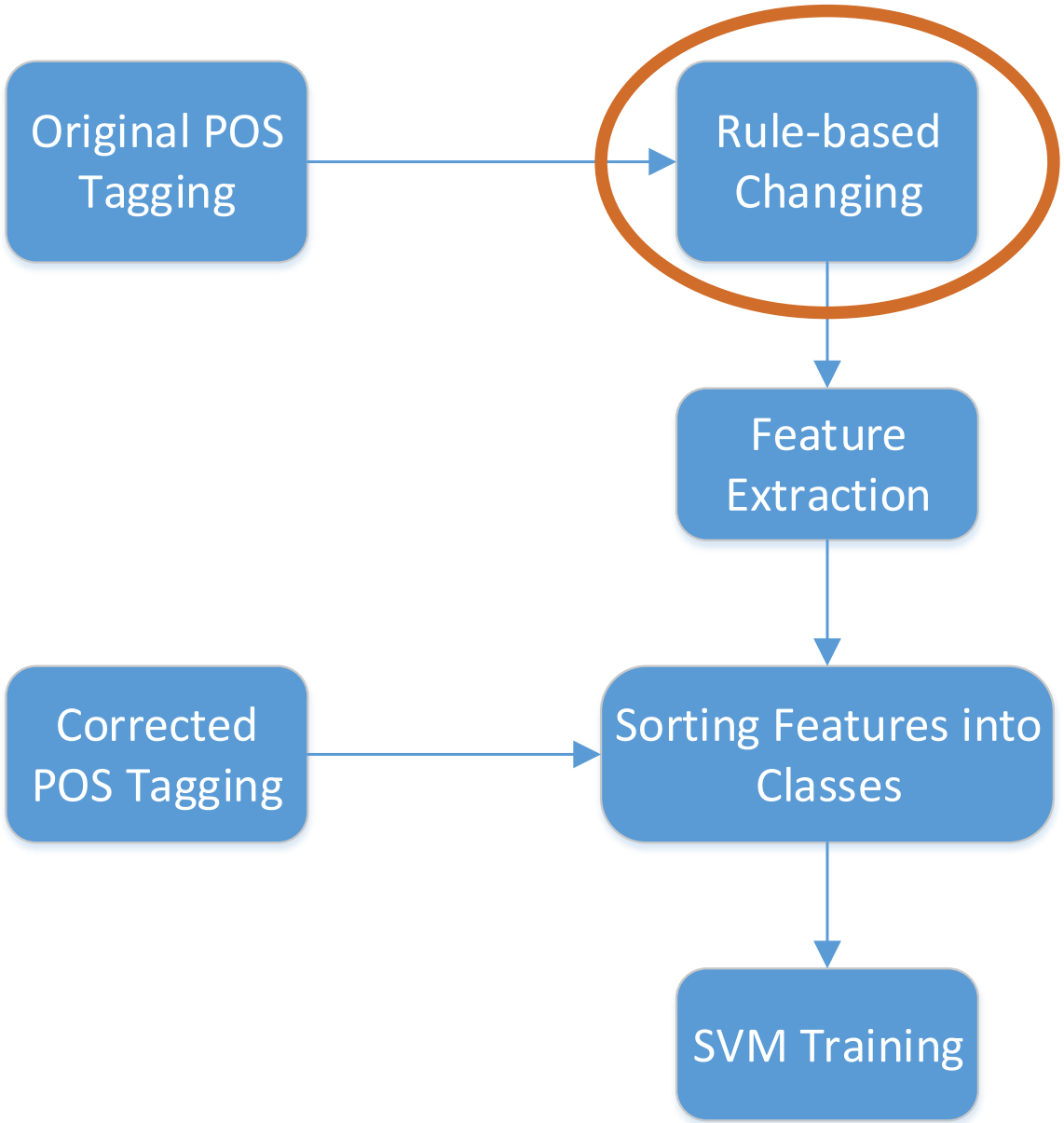
The Next Step

- Next Step after gathering all the data is to run a system to automatically correct Patent Claim POS tags.
- We needed Amazon Mechancial Turk to gather corrected tags from patent segments to serve as ground truth for developing our system

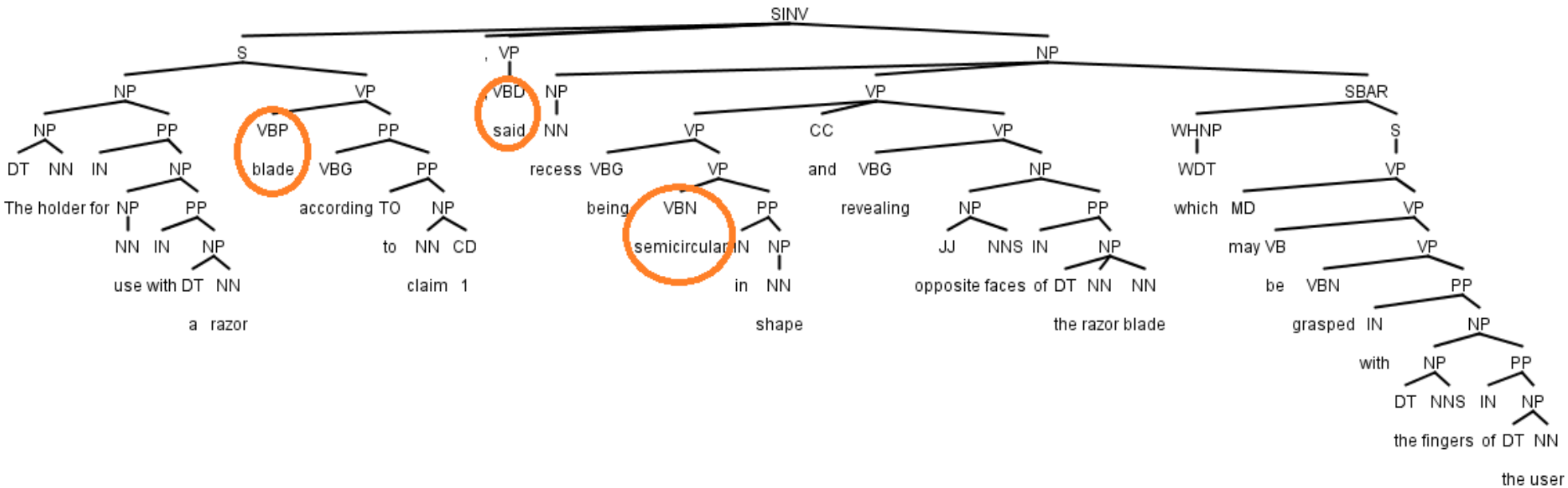
Training Stage



Rule-Based Corrector



Rule-Based Corrector

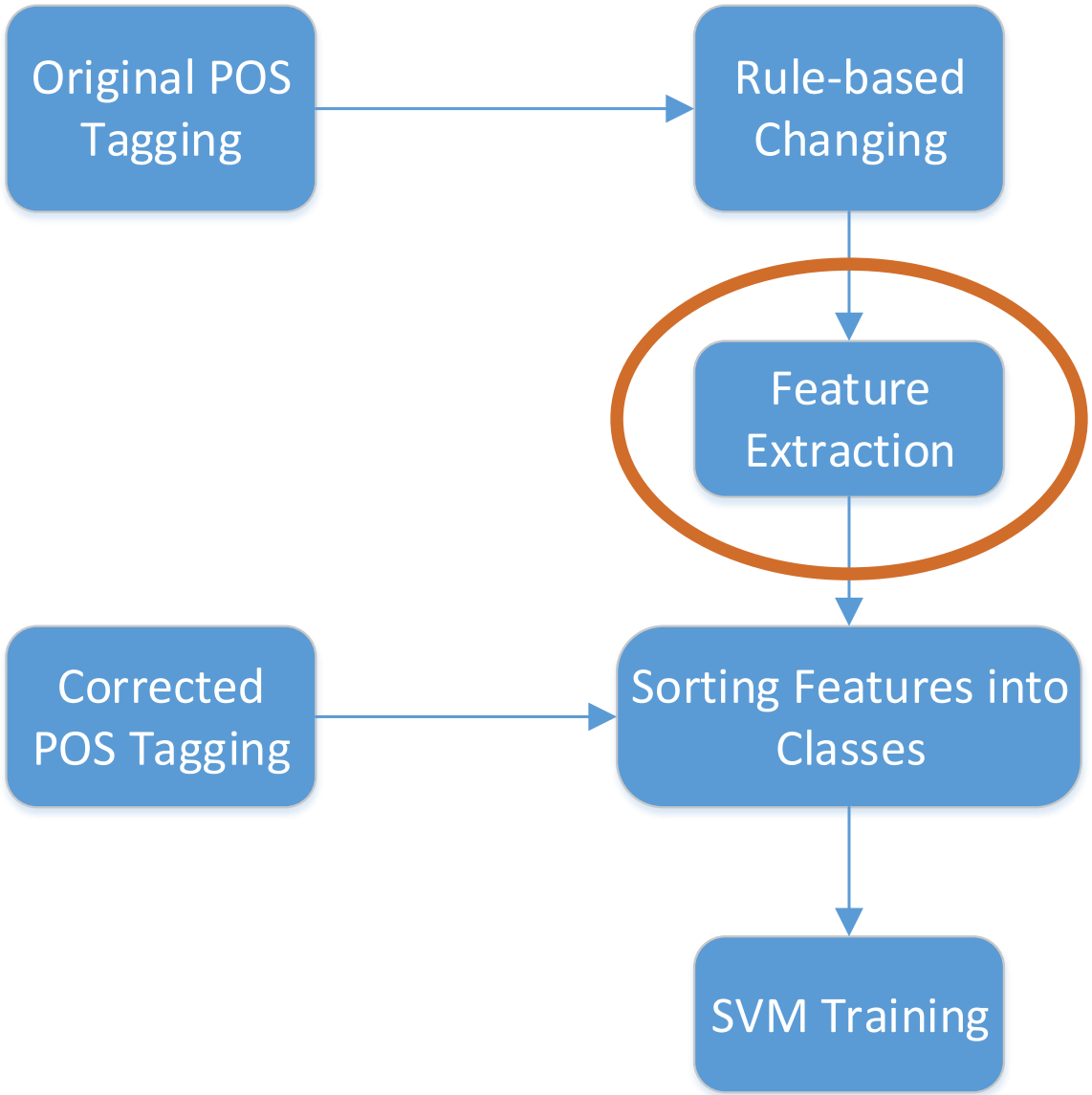


- Some words are almost always mislabeled as verbs and can thus be automatically corrected from the very beginning via a rule-like system
- List is ever expanding when new words that are always nouns or adjectives are discovered
- Will also force words in the list that have been correctly tagged to keep their tags

Rule-Based Corrector Examples

- **said** → JJ ("said recess being semicircular")
- **claim** → NN ("as recited in claim 5")
  - sole exception is at the start of a patent ("In this patent, we claim")
- **means** → NN ("including means for continuously conveying")
- **wherein** → WRB ("The system of claim 13 wherein")
- **nitride** → NN ("silver nitride")
- **boride** → NN ("cobalt boride")

Gathering Trigrams and Dependencies

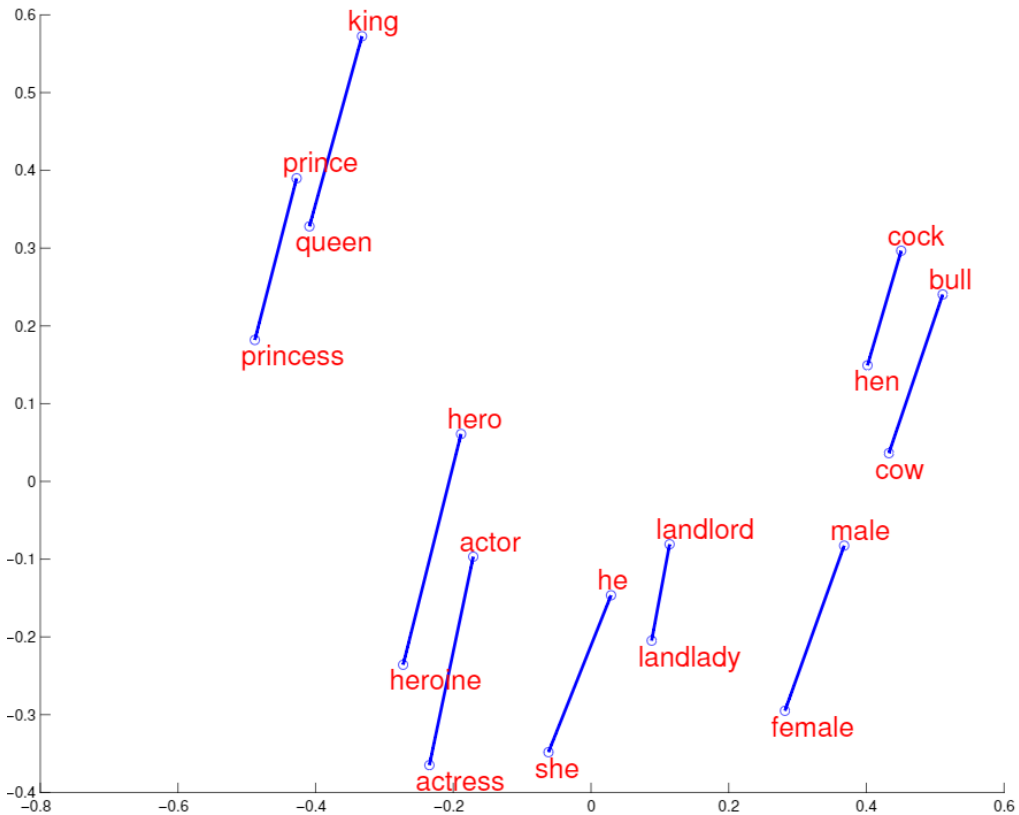


Gathering Trigrams and Dependencies

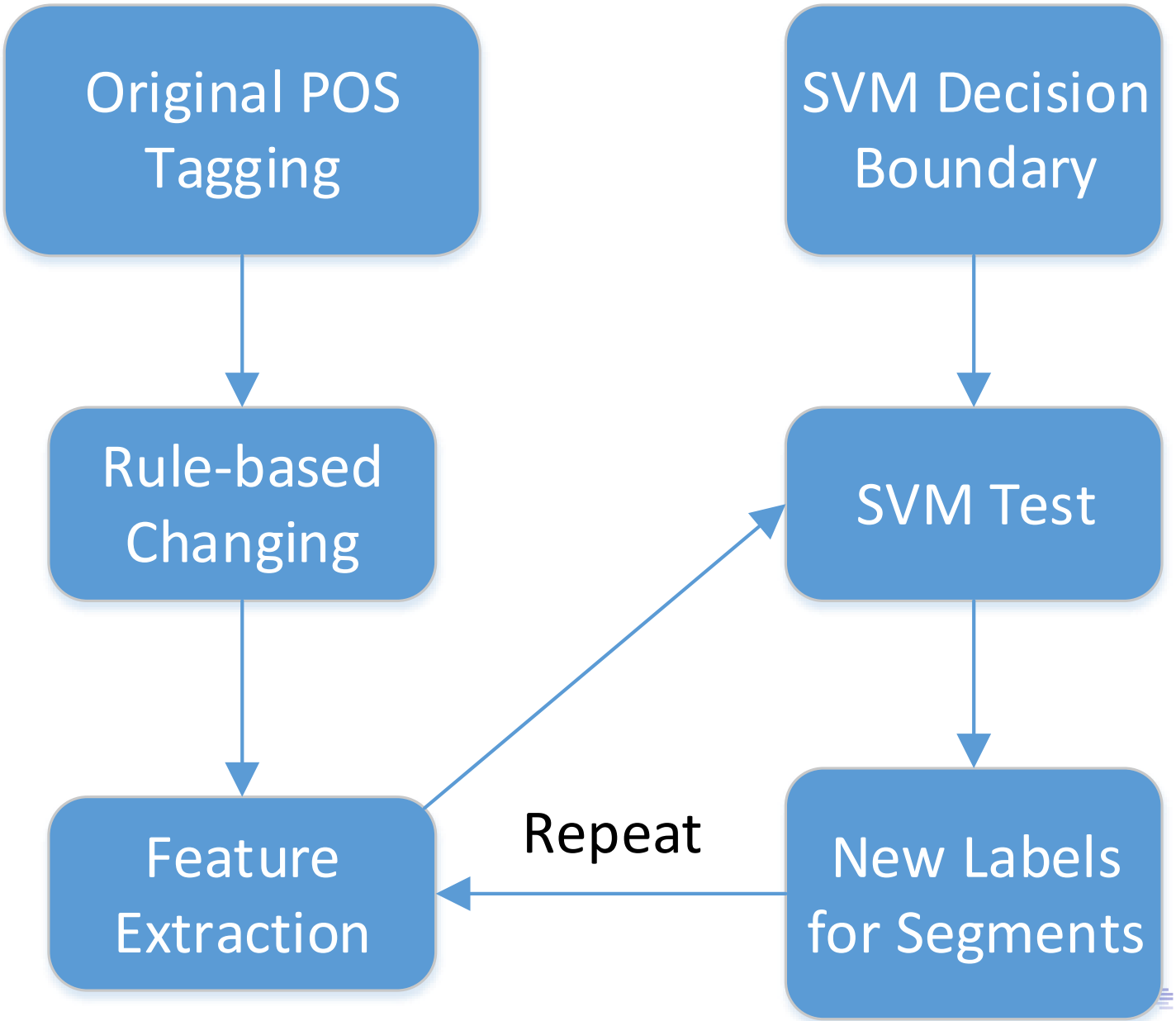
- Developed Matlab code to gather all the “verbs” and its trigrams and dependencies and filter them into whether they are actually an adjective, noun, or still a verb.
- Each trigram and set of dependencies focus on just one “verb”. The other words are represented just by their POS tags
- Question lies on how to represent the “verb” while preventing overfitting and allowing for similar words to be grouped together

Word Vectors

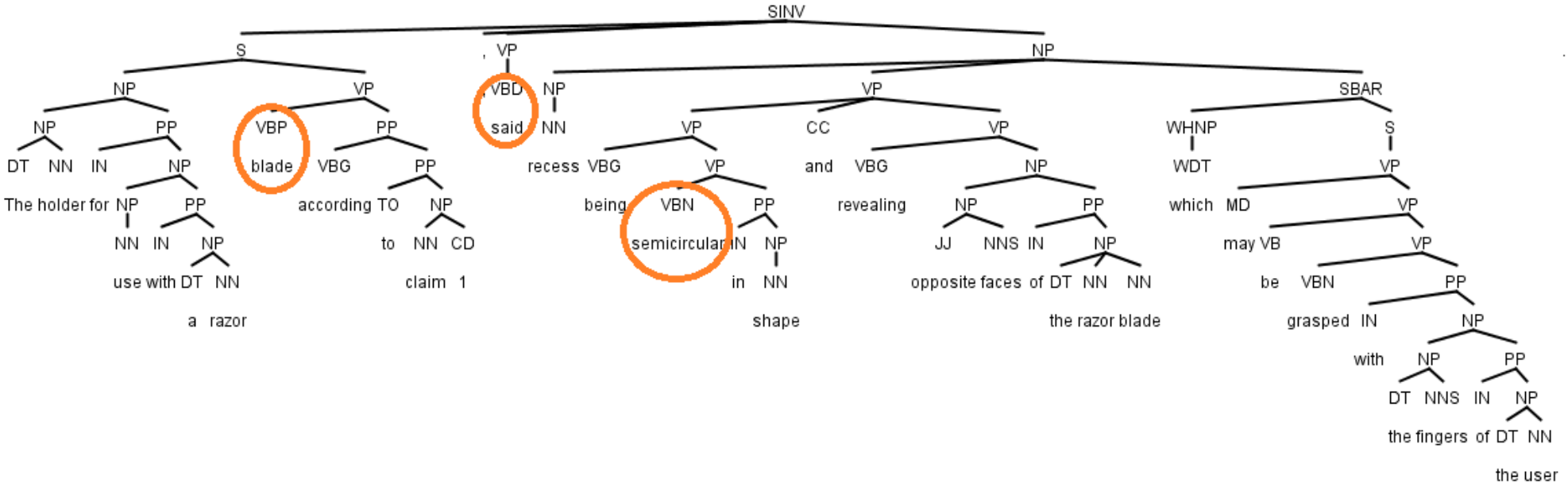
- Representation of words as a vector
- Trained on an input database (for us, patent claims)
- Preserves similar relationships between similar groups of words

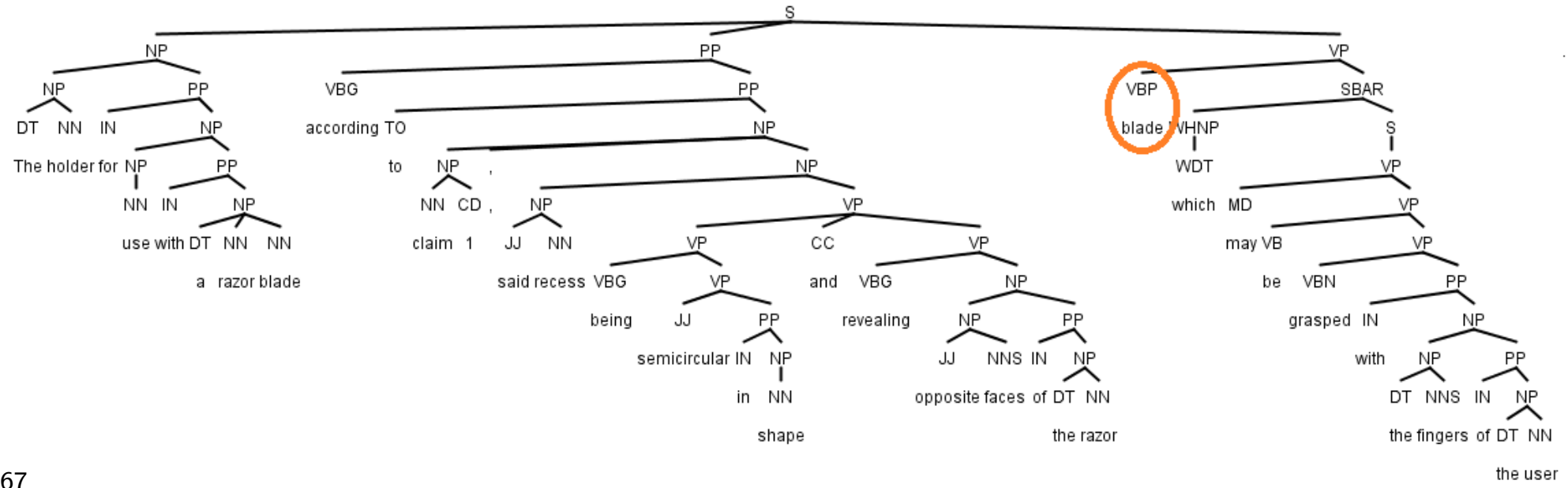


Testing Stage



Multiple Rounds of Testing?





After 4 Iterations: A Truly Corrected Parse Tree

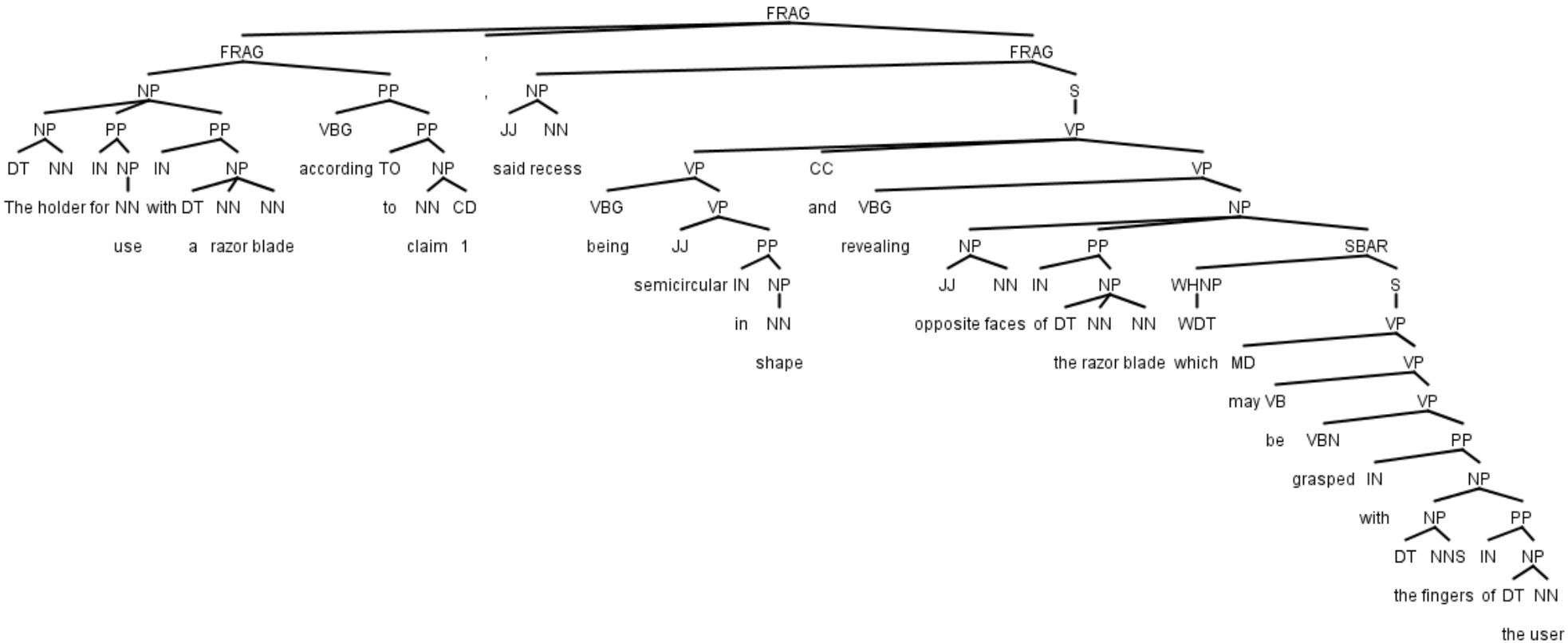


Table of Contents

Overview of Our Problem

Using Amazon Mechanical Turk
- Worker Views
- Requester Views
- Testing HITs
- Difficulties of Obtaining Results

Automatic Correcter of Patent Claim POS Tags

Conclusion

Conclusion

- Patent Claims, due to their structure, are notorously difficult for a computer to properly parse
- An automated system can be made to correct incorrect POS tags in order to be used in more advanced patent processing systems
- A large hand-corrected dataset needs to be obtained in order to learn how to properly change POS tags.
- Amazon Mechanical Turk provides the tools needed to crowdsource this hand labeling.

Further Work

- Taking the approved Amazon Mechanical Turk results, determining how correct they actually are and make a large database of the original tagging and the corrected tagging.
- Learning how to make a corpus for use in word2vec and then running word2vec on an all patent claim database.
- Putting together all the pieces of codes already developed for training and testing our automatic corrector and begin experimenting with it