Introductory Course at ESSLLI

Bolzano, Italia
August 2016

Crowdsourcing Linguistic Datasets

LECTURE 3

Chris Biemann
biem@cs.tu-darmstadt.de

Lesson 3: Quality Assurance

- Quality Control Mechanisms
- Effects of Redundancy
- Schemes and Patterns for successful data acquisition
- The HPU
- Large Sample Project

Cheap and Fast – but is it Good?

Outcry at N Korea 'nuclear test'

(Anger, 30), (Disgust,30), (Fear,30), (Joy,0), (Sadness,20), (Surprise,40), (Valence,-50).

- Systematic comparative evaluation of quality and cost of crowdworkers vs. expert annotators
- several tasks, example (on right) for "affective text analysis"

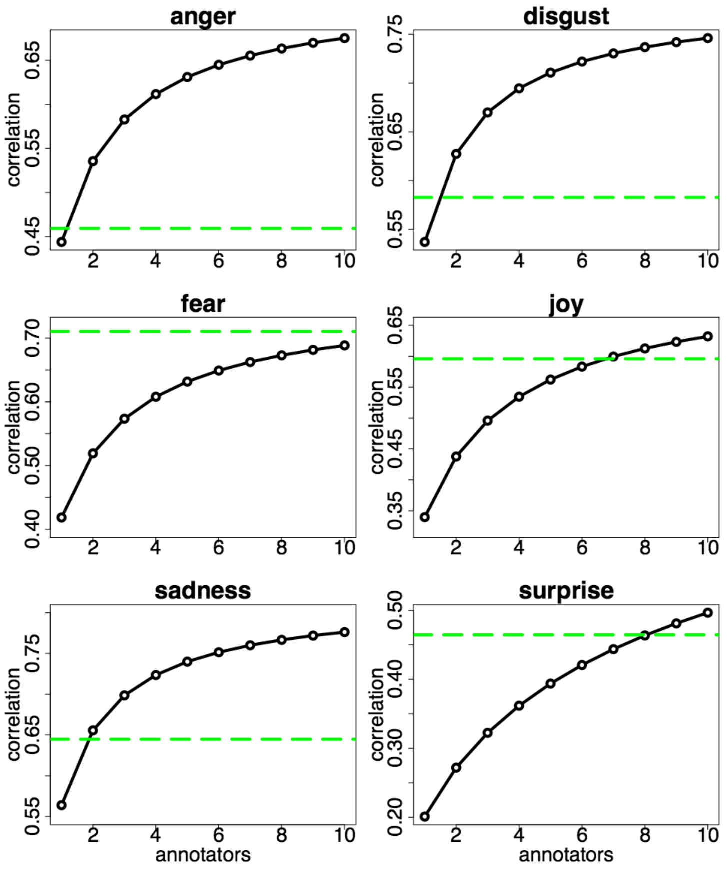


Figure 1: Non-expert correlation for affect recognition

Main takeaways:

- experts are better but more expensive
- average of 4 crowdworkers needed to replace 1 expert
- controlling for label bias and quality helps

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-08). pp. 254-263.

Quality Control Mechanisms

- Reputation systems: stats about approval rate, hits completed etc.
- Qualification tests: user-defined tests
- Aggregation and redundancy: don’t trust a single worker, trust the crowd!
- Embedded gold standard data: trust is good, checking is better
- Second-pass reviewing: turk the turkers
- Economic incentives: bonus for agreement
- Statistical models: embrace the noise and attribute more weight to reliable crowdworkers

Behavioral Patterns of Malicious Workers

Ineligible Workers (IW)

Instruction: Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.

Response: *'this is my first task'*

Fast Deceivers (FD)

eg: Copy-pasting same text in response to multiple questions, entering gibberish, etc.

Response: *'What's your task?'* , *'adasd'*, *'fgfgf gsd ljlkj'*

Rule Breakers (RB)

Instruction: Identify 5 keywords that represent this task (separated by commas).

Response: *'survey, tasks, history'* , *'previous task yellow'*

Smart Deceivers (SD)

Instruction: Identify 5 keywords that represent this task (separated by commas).

Response: *'one, two, three, four, five'*

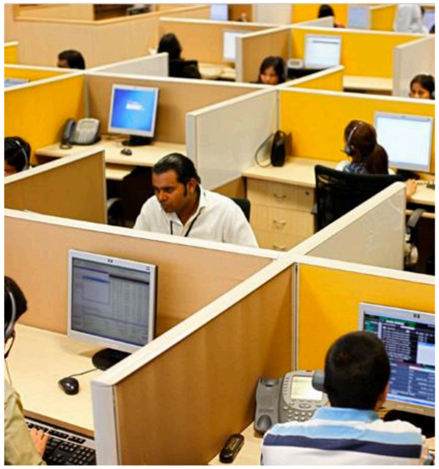
Gold Standard Preys (GSP)

These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions!

Gadiraju, U., Kawase, R., Dietze, S., Demartini, G. (2015): Understanding Malicious Behaviour in Crowdsourcing Platforms: The Case of Online Surveys. In: Proceedings of the ACM Special Interest Group on Computer Human Interaction (CHI 2015)

Indian Spam

(2013) While there are certainly many high-quality workers outside the US, there is a certain segment of workers that join the market with the **sole purpose of getting something for nothing**. Especially after Indian workers became eligible to receive cash compensation (instead of just gift cards available to other non-US workers), the **number of spam attacks from India** went up significantly.

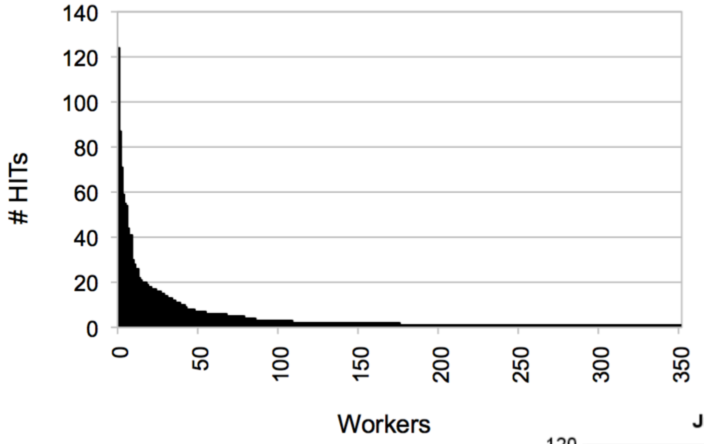


Business process outsourcing $1/hr

Rural work $0.25/hr

Iperotis, P. (2013): Mechanical Turk account verification: Why Amazon disables so many accounts. http://www.behind-the-enemy-lines.com/2013/06/mechanical-turk-account-verification.html
https://rr.soe.ucsc.edu/sites/default/files/2010-davis.pdf

Typical Distributions of HITs per Crowdworker



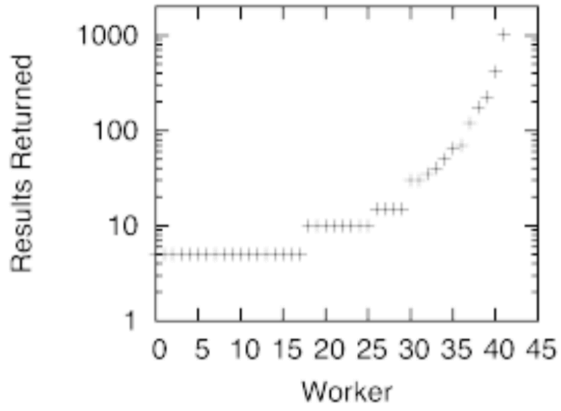
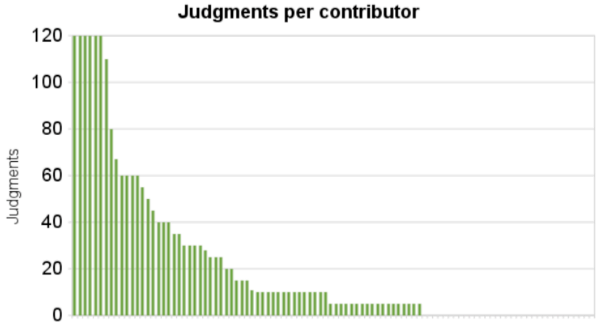


Figure 4: Only about one-third of the workers did more than three HITs and a a few prolific workers accounted for most of our data.



- Just like anything else that is uncontrolled and unbound, distribution follows a power law: ~half of the people do only 1 HIT
- Quality and number of HITs does not correlate, but there are typical ‘outliers’: high-volume scammers, low-volume tryouts...

$300 project: Acquire complex paraphrases

- Previous project: paraphrase candidates given
- reached agreement through formalizing/constraining the interface
- How to reach agreement for **collection** of paraphrases?
  - high variability of correct answers
  - thus, no ‘gold’ items
  - how to reach high quality?
- Solution: use crowdworkers for validation!

Tschirsich, M. and Hintz, G. (2013): Leveraging Crowdsourcing for Paraphrase Recognition. Proceedings of LAW and Interoperability with Discourse, Sofia, Bulgaria, pp. 205--213

Paraphrase generation

Crowdflower task design

- Short text chunks
  - Verb phrases (most promising for paraphrasing)
  - 3 to 8 words
  - Extracted with NLTK chunking grammar

```
NP: {<DT|JJ.*|CD|PRP\$>*<NN.*>+}
PP: {<IN><NP>}
Arg: {<NP|PP>}
Verb: {<MD>?<VB.*>+<PRP|R.*>?}
VerbPhrase: {<Verb><Arg>}
```

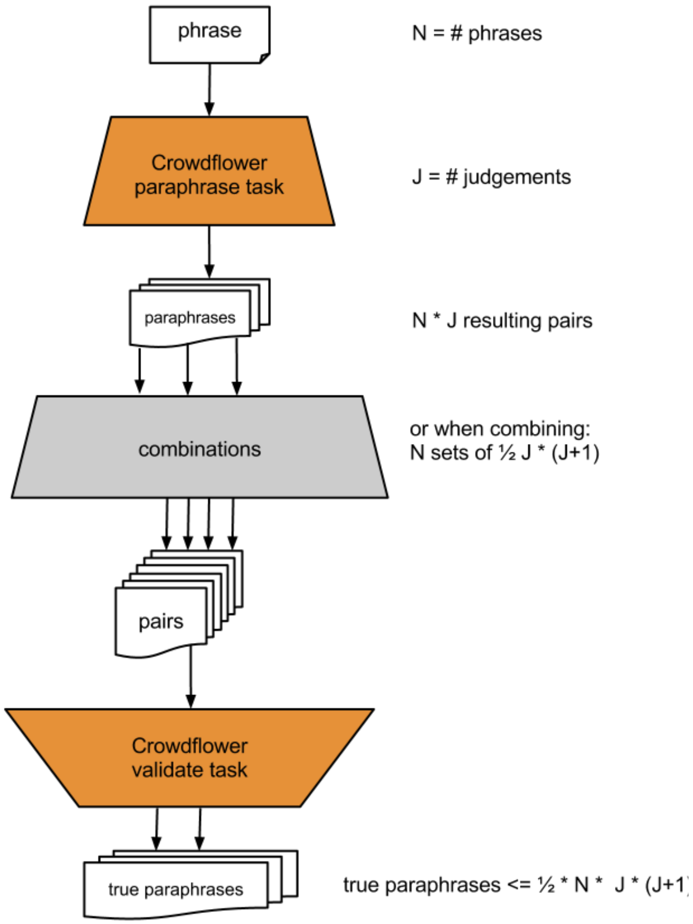
- Source corpus: MSRP (Microsoft Research Paraphrase Corpus)
- HIT: “Please paraphrase the given expression”
- Prevent bad input as good as possible

Two/Multi-Stage Generation

Simple multi-stage generation process as proposed by Negri et. al, 2012

1. Text chunks to paraphrase
2. Obtain crowd paraphrases
3. (Optional) further iterations to increase lexical divergence
4. Use a validation task to filter bad data

- The validate step can have gold
- But generation step is just chance



phrase

N = # phrases

Crowdflower paraphrase task

J = # judgements

paraphrases

N * J resulting pairs

combinations

or when combining:
N sets of ½ J * (J+1)

pairs

Crowdflower validate task

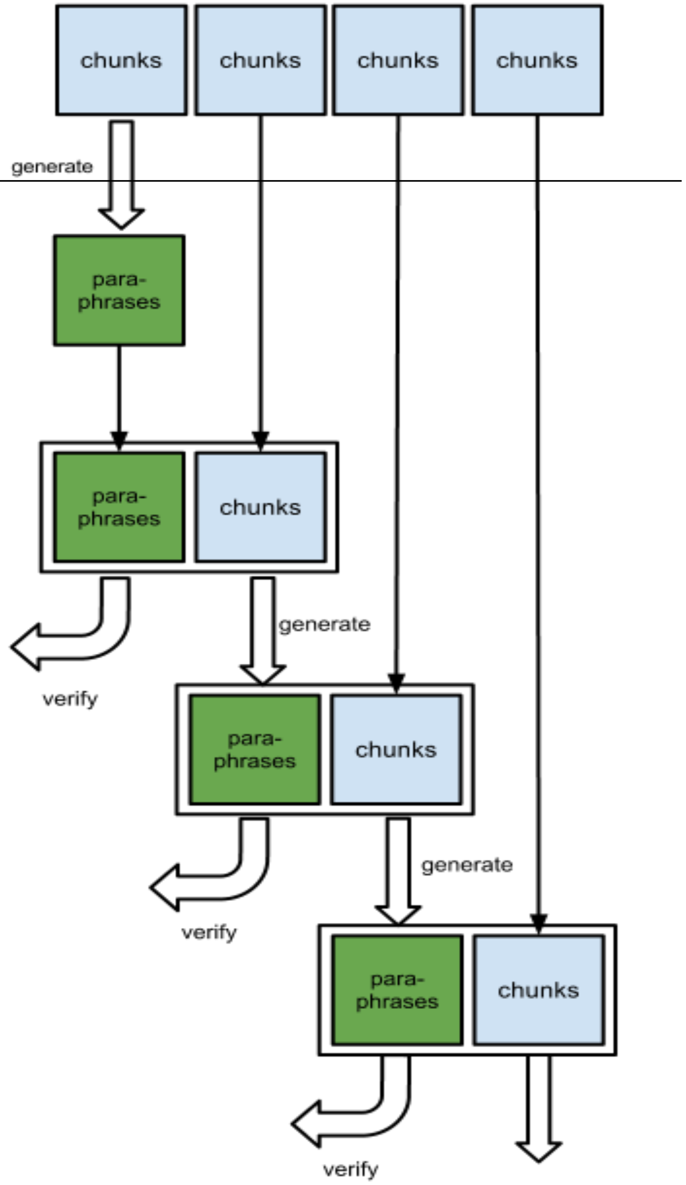
true paraphrases

true paraphrases <= ½ * N * J * (J+1

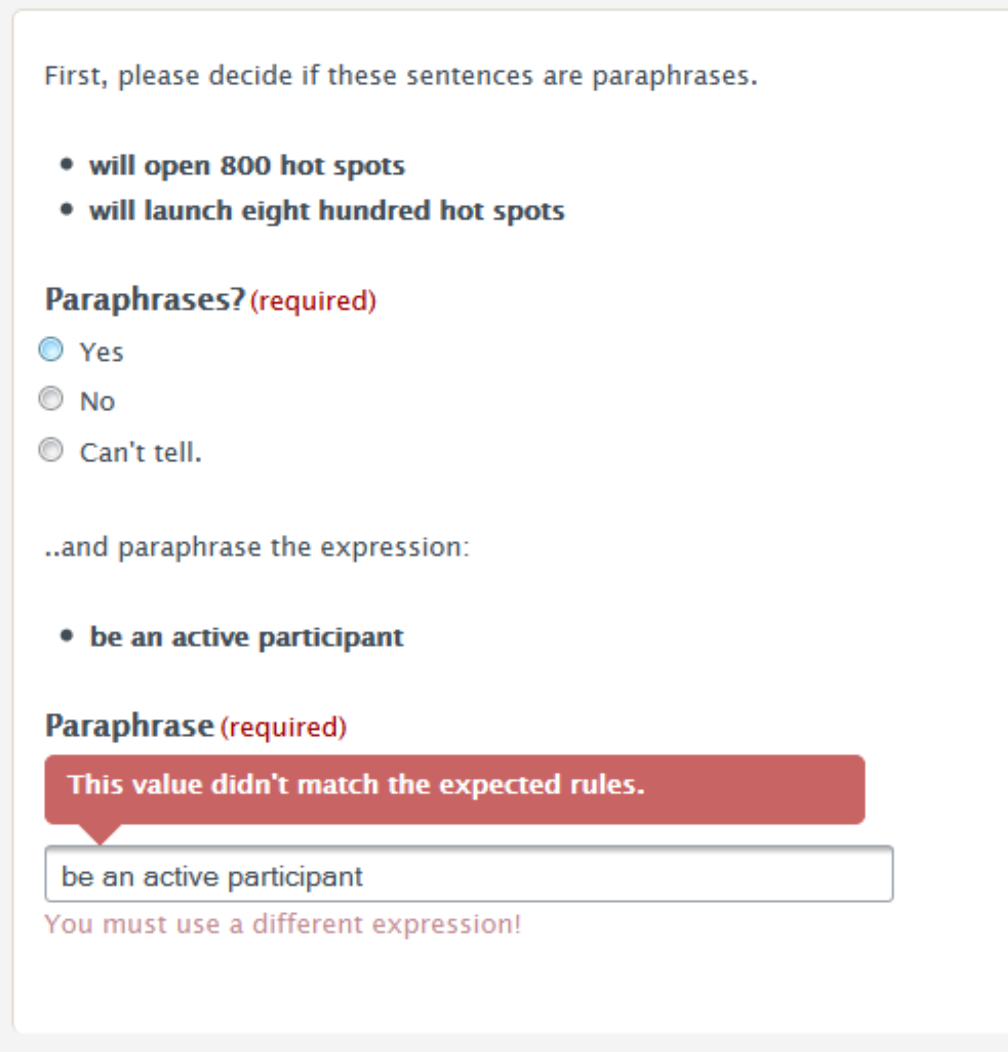
Negri, M., Mehdad, Y., Marchetti, A., Giampiccolo, D., and Bentivogli, L. (2012): Chinese whispers: Cooperative paraphrase acquisition. In Proceedings of LREC'12, Istanbul, Turkey

Combined Task Scheme

- Combine generate / verify HITS into one atomic unit
- Initial generate
- Pair each generate HIT with validate HIT
- Use gold on validate to *infer* generate quality



Combined Task Interface (Crowdflower)



First, please decide if these sentences are paraphrases.

- will open 800 hot spots
- will launch eight hundred hot spots

Paraphrases? (required)

○ Yes

○ No

○ Can't tell.

..and paraphrase the expression:

- be an active participant

Paraphrase (required)

This value didn't match the expected rules.

be an active participant

You must use a different expression!

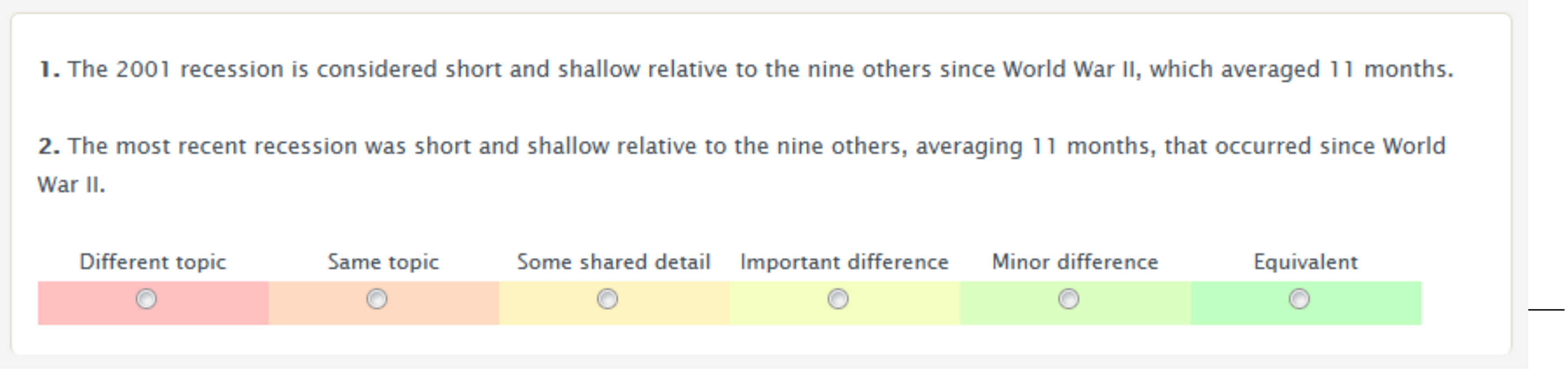
- In the combined setting, spamming on half of the HIT is still possible, but rather unlikely
- use JQuery rules to avoid copy-and-paste of the expression as well as empty text
- bad data caught early, bad workers can be blocked

Paraphrase validation

- Two possibilities
  - **Binary verification**
    ask the worker to make a binary decision Yes / No
  - **Semantic similarity rating**
    ask the worker to rate the similarity

Choice depends on goal:

- binary is easier for agreement
- n-ary needs less judgments for converging graded score, but costs more



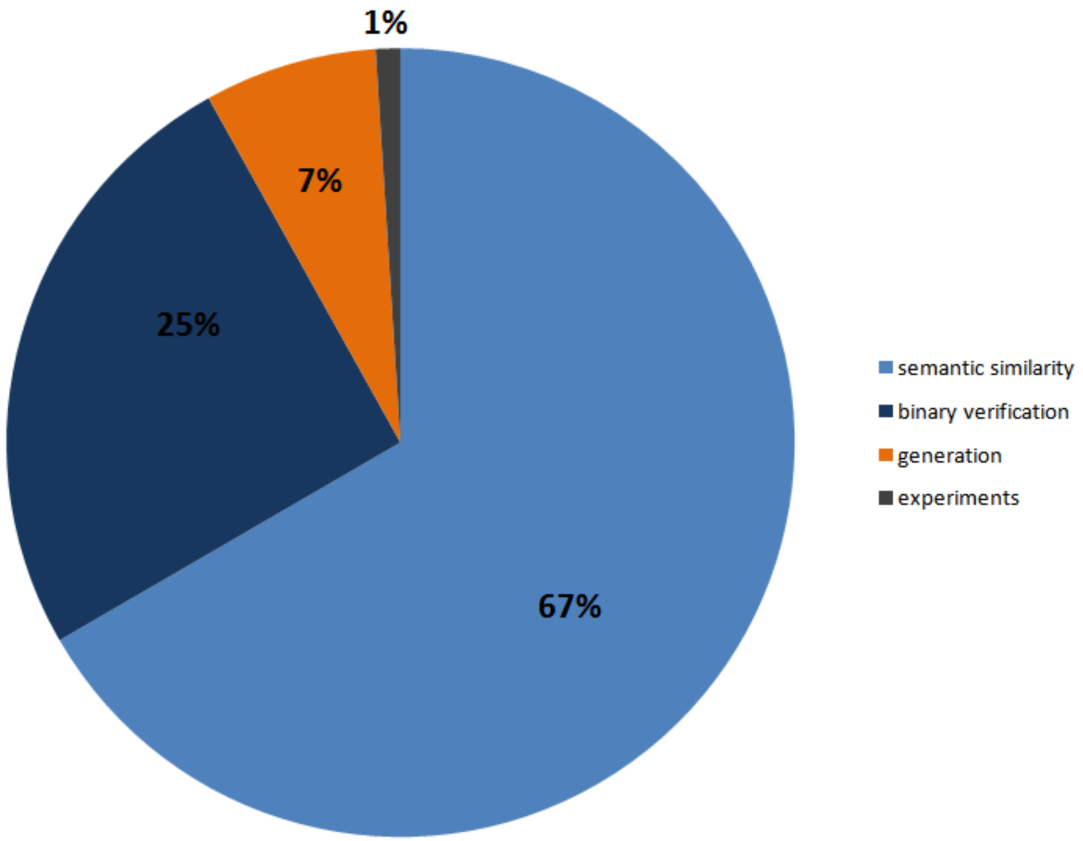
1. The 2001 recession is considered short and shallow relative to the nine others since World War II, which averaged 11 months.

2. The most recent recession was short and shallow relative to the nine others, averaging 11 months, that occurred since World War II.

Different topic	Same topic	Some shared detail	Important difference	Minor difference	Equivalent
○	○	○	○	○	○

Complex Paraphrase Generation: Task investment

- Total funds $300
  - 9 different tasks
  - 4684 units
  - 16552 judgments
  - 1.8 ¢ / judgment



- Focus on verification
  - Binary
  - Semantic Similarity

Let’s Crowdsource:
Align Definitions by Meaning

WordNet: CONE#N

- 1. (n) cone (any cone-shaped artifact)
- 2. (n) cone, conoid, cone shape (a shape whose base is a circle and whose sides taper up to a point)
- 3. (n) cone, strobilus, strobile (cone-shaped mass of ovule- or spore-bearing scales or bracts)
- 4. (n) cone, cone cell, retinal cone (a visual receptor cell in the retina that is sensitive to bright light and to color)

Wikionary: CONE#N

- 1. A surface of revolution formed by rotating a segment of a line around another line that intersects the first line.
- 2. A solid of revolution formed by rotating a triangle around one of its altitudes.
- 3. A space formed by taking the direct product of a given space with a closed interval and identifying all of one end
- to a point.
- 4. Anything shaped like a cone.
- 5. The fruit of a conifer.
- 6. An ice cream cone.
- 7. A unit of volume, applied solely to marijuana and only while it is in a smokable state; roughly 1.5 cubic centimetres, depending on use.
- 8. Any of the small cone-shaped structures in the retina.

Let’s Crowdsource!

Now, How about adding these?

Urban dictionary: CONE#N

1. Metallic cone shaped item in a bong or pipe that the weed is burnt in. Another word for a ’bowl’.
2. In memphis CONE means a junky or person who is always begging or asking for something or even a lame person
3. A cone is the bowl of a bong where the weed is put.
4. Cone-shaped life form, often orange but occasionally green and very rarely fluoro yellow. Often gathers with similar species around construction sites. Different species of cones are rarely found intermixed with one another.
5. A joint of marijuana rolled on an angle making it a cone shaped
6. A passenger on a cruise ship.
7. a social reject. one who does not fit in, or is considered repulsive by others

WordNet: CONE#N

1. (n) cone (any cone-shaped artifact)
2. (n) cone, conoid, cone shape (a shape whose base is a circle and whose sides taper up to a point)
3. (n) cone, strobilus, strobile (cone-shaped mass of ovule- or spore-bearing scales or bracts)
4. (n) cone, cone cell, retinal cone (a visual receptor cell in the retina that is sensitive to bright light and to color)

Wikionary: CONE#N

1. A surface of revolution formed by rotating a segment of a line around another line that intersects the first line.
2. A solid of revolution formed by rotating a triangle around one of its altitudes.
3. A space formed by taking the direct product of a given space with a closed interval and identifying all of one end to a point.
4. Anything shaped like a cone.
5. The fruit of a conifer.
6. An ice cream cone.
7. A unit of volume, applied solely to marijuana and only while it is in a smokable state; roughly 1.5 cubic centimetres, depending on use.
8. Any of the small cone-shaped structures in the retina.

Design Pattern: Find-Fix-Verify

Find-Fix-Verify splits complex crowd intelligence tasks into a series of generation and review stages that utilize independent agreement and voting to produce reliable results.

Separate open-ended tasks into three stages:

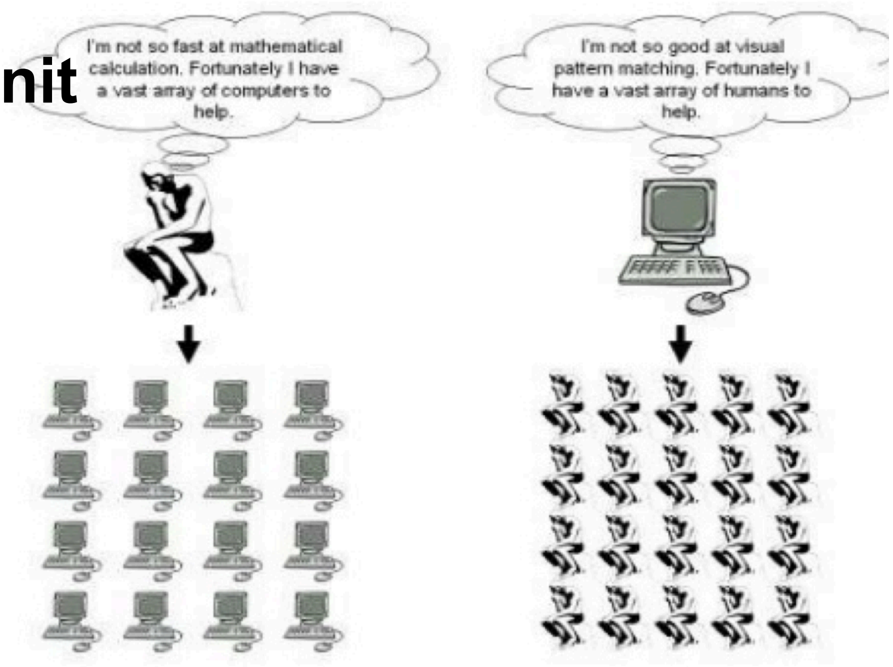
- Find: Identify areas that need work / errors
  aggregate to determine what should be fixed
- Fix: Propose solutions to identified problems
  collect only a few alternatives
- Verify: Vote on the alternatives to find the best one

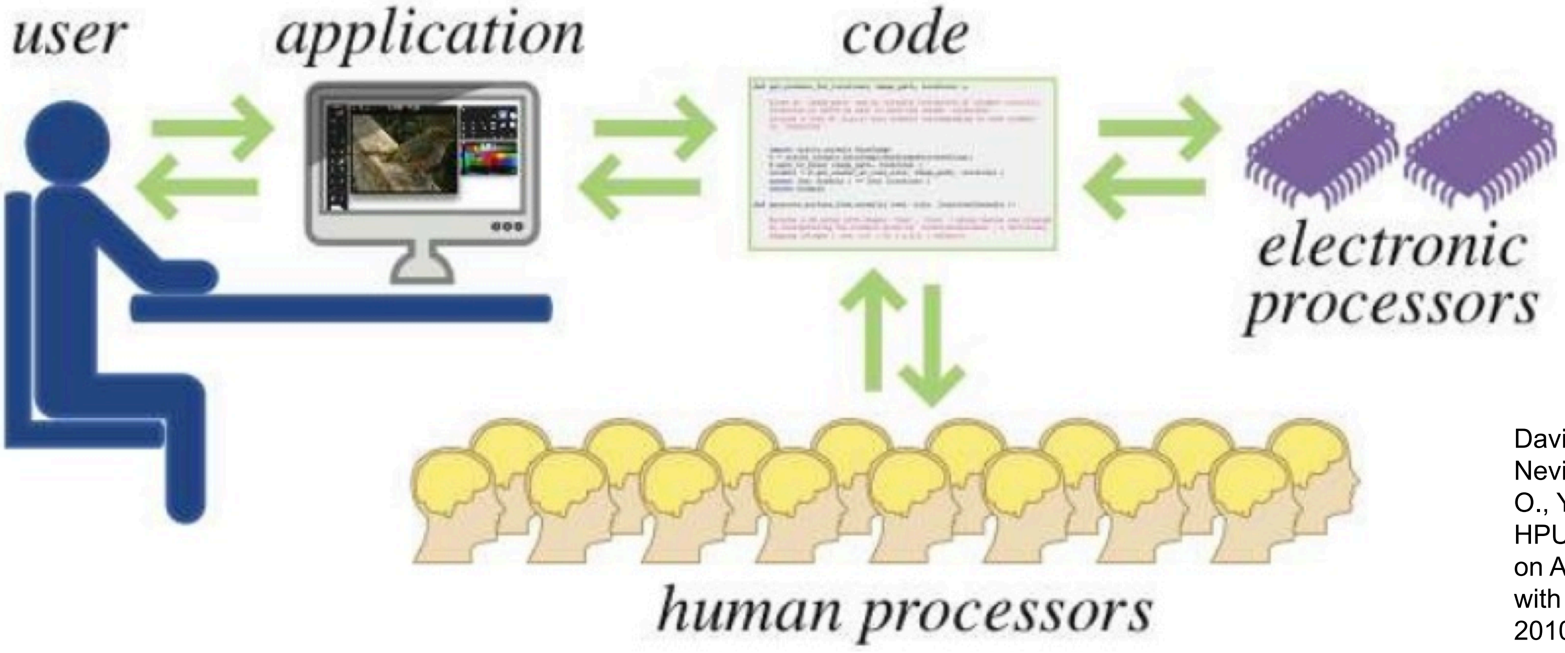
Typically, enforce different workforce on the three stages

Bernstein, M., Little, G., Miller, R.C., Hartmann, B., Ackerman, M., Karger, D.R., Crowell, D., and Panovich, K. (2010): Soylent: A Word Processor with a Crowd Inside. In Proc. UIST 2010. ACM Press.

The HPU = Human Processing Unit

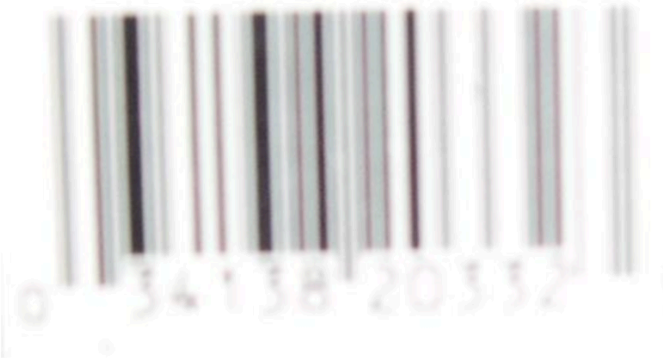
- Humans as a first-class computational platform
- Can execute functions on HPUs that return values

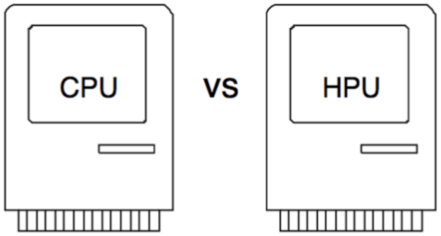

I'm not so fast at mathematical calculation. Fortunately I have a vast array of computers to help.
I'm not so good at visual pattern matching. Fortunately I have a vast array of humans to help.

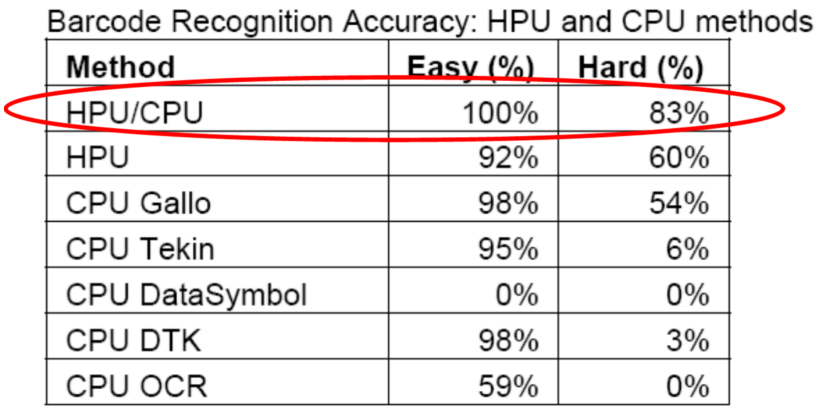


Davis, J., Arderiu, J., Lin, H., Nevins, Z., Schuon, S., Gallo, O., Yang, M.-H. (2010): The HPU. IEEE CVPR Workshop on Advancing Computer Vision with Humans in the Loop, 2010.

Example: Accuracy of Bar Codes with CPUs and HPUs







Barcode Recognition Accuracy: HPU and CPU methods

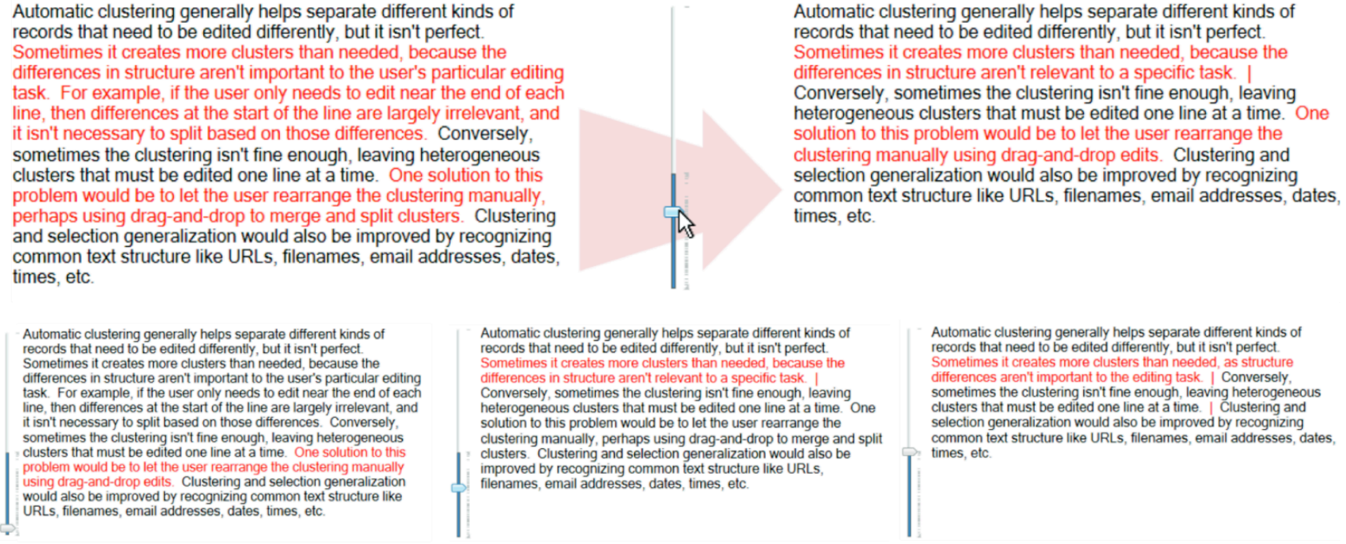
Method	Easy (%)	Hard (%)
HPU/CPU	100%	83%
HPU	92%	60%
CPU Gallo	98%	54%
CPU Tekin	95%	6%
CPU DataSymbol	0%	0%
CPU DTK	98%	3%
CPU OCR	59%	0%

- if combined in a good way, CPUs and HPUs help each other

https://rr.soe.ucsc.edu/sites/default/files/2010-davis.pdf

Soylent: A Word Processor With a Crowd Inside

http://projects.csail.mit.edu/soylent/



Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't relevant to a specific task. | Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

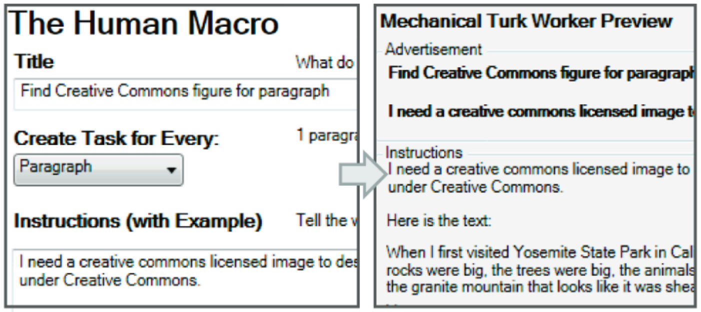
Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't relevant to a specific task. | Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, as structure differences aren't important to the editing task. | Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. | Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Figure 1. Shortn allows users to adjust the length of a paragraph via a slider. Red text indicates locations where cuts or rewrites have occurred. Tick marks represent possible lengths, and the blue background bounds the possible lengths.

- Available as an Open Source project
- Check out their Video!
- Claimed error tolerance: 30%



The Human Macro

Title

What do

Find Creative Commons figure for paragraph

Create Task for Every:

1 paragr

Paragraph

Instructions (with Example)

Tell the v

I need a creative commons licensed image to des under Creative Commons.

Mechanical Turk Worker Preview

Advertisement

Find Creative Commons figure for paragraph

I need a creative commons licensed image t

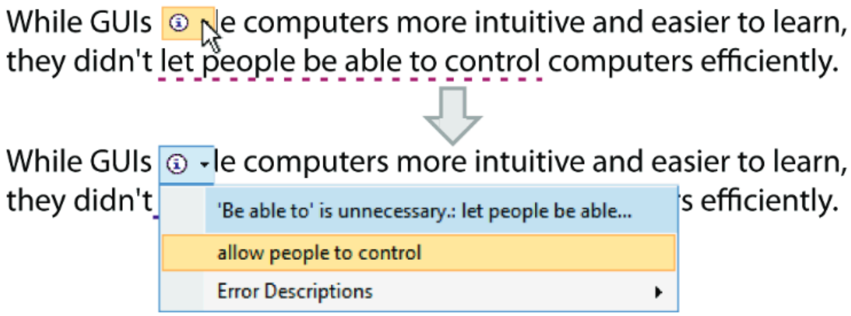
Instructions

I need a creative commons licensed image to under Creative Commons.

Here is the text:

When I first visited Yosemite State Park in Cal rocks were big, the trees were big, the animals the granite mountain that looks like it was shea

Figure 3. The Human Macro is an end-user programming interface for automating document manipulations. The left half is the user's authoring interface; the right half is a preview of what the Turker will see.



While GUIs ⓘ le computers more intuitive and easier to learn, they didn't let people be able to control computers efficiently.

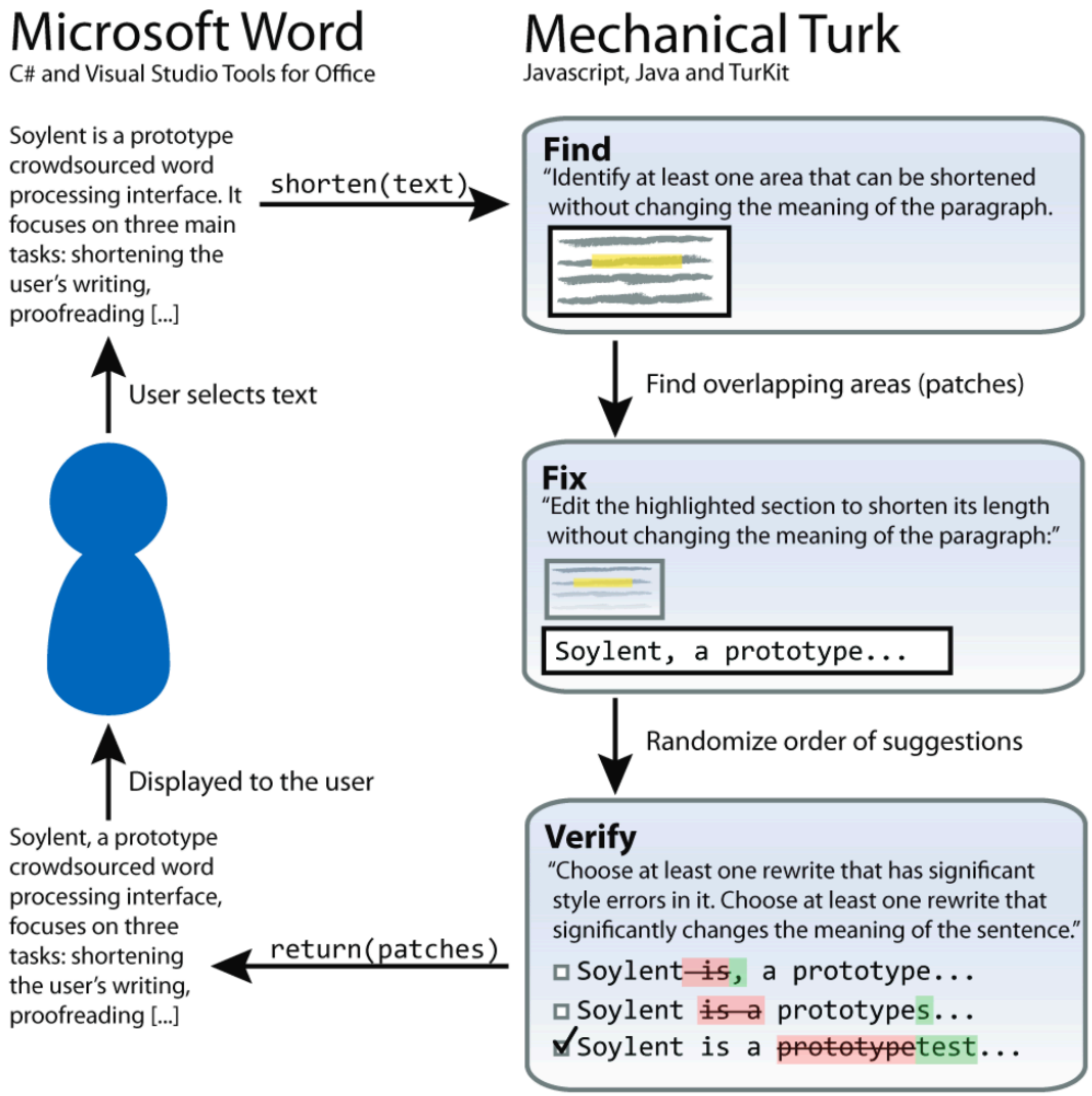
While GUIs ⓘ le computers more intuitive and easier to learn, they didn't s efficiently.

Figure 2. Crowdproof is a human-augmented proofreader. The drop-down explains the problem (blue title) and suggests fixes (gold selection).

Find-Fix-Verify in Soylent

Shortn experiment:

- pricing: $0.08 per Find, $0.05 per Fix, $0.04 per Verify
- Average paragraph cost $1.41:
  - $0.55 to identify avg. of two patches,
  - $0.48 to generate alternatives and
  - $0.38 to filter results



Microsoft Word

C# and Visual Studio Tools for Office

Mechanical Turk

Javascript, Java and TurKit

Soylent is a prototype crowdsourced word processing interface. It focuses on three main tasks: shortening the user's writing, proofreading [...]

`shorten(text)`

Find

"Identify at least one area that can be shortened without changing the meaning of the paragraph.

User selects text

Find overlapping areas (patches)

Fix

"Edit the highlighted section to shorten its length without changing the meaning of the paragraph:"

`Soylent, a prototype...`

Randomize order of suggestions

Displayed to the user

Verify

"Choose at least one rewrite that has significant style errors in it. Choose at least one rewrite that significantly changes the meaning of the sentence."

□ `Soylent ~~is~~, a prototype...`

□ `Soylent ~~is a~~ prototypes...`

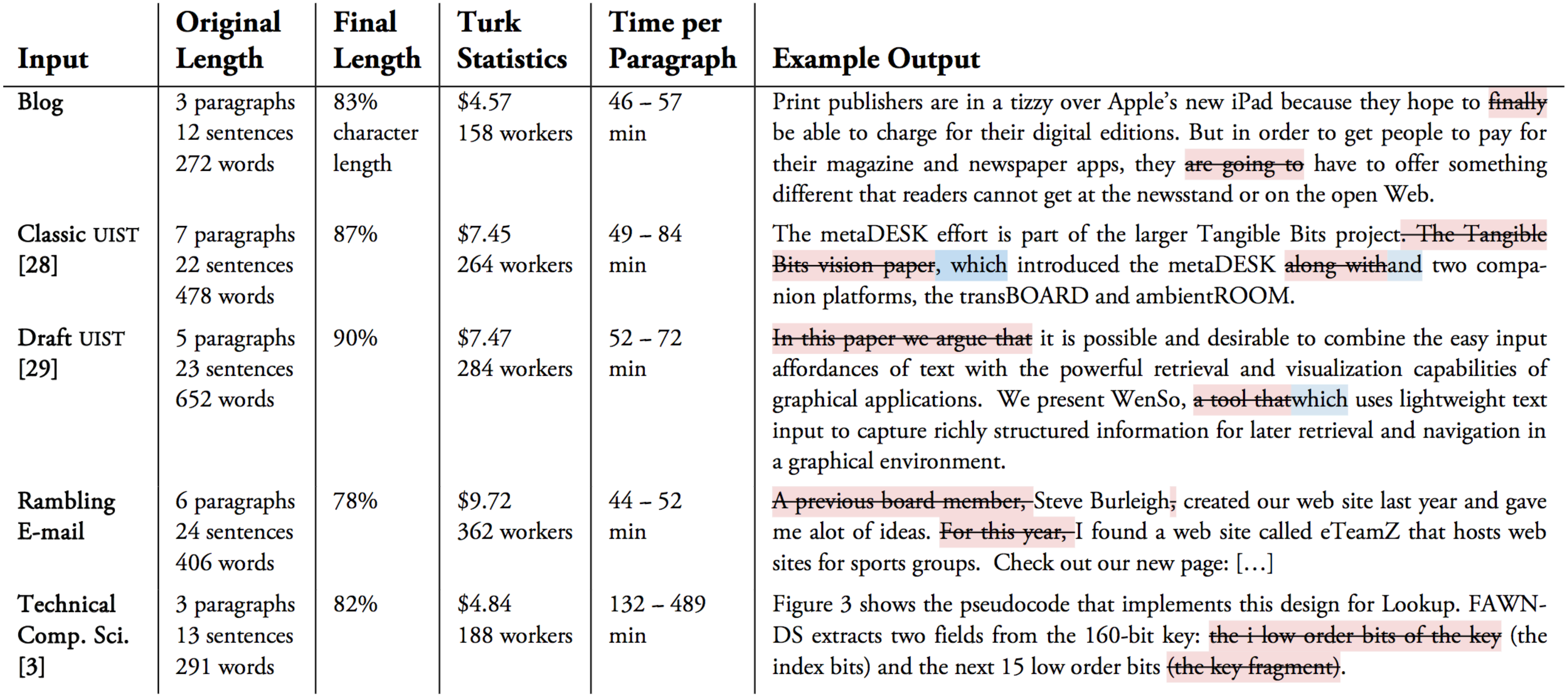
☑ `Soylent is a ~~prototype~~test...`

Soylent, a prototype crowdsourced word processing interface, focuses on three tasks: shortening the user's writing, proofreading [...]

`return(patches)`

Figure 4. Find-Fix-Verify identifies patches in need of editing, recruits workers to fix the patches, and votes to approve work.

Soylent Shortn example



Input	Original Length	Final Length	Turk Statistics	Time per Paragraph	Example Output
Blog	3 paragraphs 12 sentences 272 words	83% character length	$4.57 158 workers	46 – 57 min	Print publishers are in a tizzy over Apple's new iPad because they hope to ~~finally~~ be able to charge for their digital editions. But in order to get people to pay for their magazine and newspaper apps, they ~~are going to~~ have to offer something different that readers cannot get at the newsstand or on the open Web.
Classic UIST [28]	7 paragraphs 22 sentences 478 words	87%	$7.45 264 workers	49 – 84 min	The metaDESK effort is part of the larger Tangible Bits project~~. The Tangible Bits vision paper~~, which introduced the metaDESK ~~along with~~and two companion platforms, the transBOARD and ambientROOM.
Draft UIST [29]	5 paragraphs 23 sentences 652 words	90%	$7.47 284 workers	52 – 72 min	~~In this paper we argue that~~ it is possible and desirable to combine the easy input affordances of text with the powerful retrieval and visualization capabilities of graphical applications. We present WenSo, ~~a tool that~~which uses lightweight text input to capture richly structured information for later retrieval and navigation in a graphical environment.
Rambling E-mail	6 paragraphs 24 sentences 406 words	78%	$9.72 362 workers	44 – 52 min	~~A previous board member,~~ Steve Burleigh~~,~~ created our web site last year and gave me alot of ideas. ~~For this year,~~ I found a web site called eTeamZ that hosts web sites for sports groups. Check out our new page: […]
Technical Comp. Sci. [3]	3 paragraphs 13 sentences 291 words	82%	$4.84 188 workers	132 – 489 min	Figure 3 shows the pseudocode that implements this design for Lookup. FAWN-DS extracts two fields from the 160-bit key: ~~the i low order bits of the key~~ (the index bits) and the next 15 low order bits ~~(the key fragment)~~.

Table I. Our evaluation run of Shortn produced revisions between 78% – 90% of the original paragraph length on a single run. The Example Output column contains example edits from each input.

Find-Fix-Verify in Machine Translation

- Find: show automatically translated text and identify incorrect grammar or incorrect translation
- Fix: ask to re-translate the wrong ones
- Verify: select the best translation

Other Hybrid (HPU) Applications

Hybrid Applications connect automation and human computation to achieve what none of their parts can do on their own.

Further Examples:

- S. Cooper et al. (2010): Predicting protein structures with a multiplayer online game
- C. Hu et al. (2010): Translation by iterative collaboration between monolingual users
- T. Yan et al. (2010): CrowdSearch: Exploiting Crowds for Accurate Real-Time Image Search on Mobile Phones

Cooper S., Khatib F., Treuille A., Barbero J., Lee J., Beenen M., Leaver-Fay A., Baker D., Popović Z., Players F. (2010): Predicting protein structures with a multiplayer online game. Nature. Aug 5;466(7307):756-60

Hu, C., Bederson, B.B. and Resnik, P. (2010): Translation by iterative collaboration between monolingual users. In Proceedings of Graphics Interface 2010 (GI '10). Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 39-46.

Yan, T., Kumar, V., Ganesan, D. (2010): CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. In Proceedings of the 8th international conference on Mobile systems, applications, and services (MobiSys '10). ACM, New York, NY, USA, 77-90.

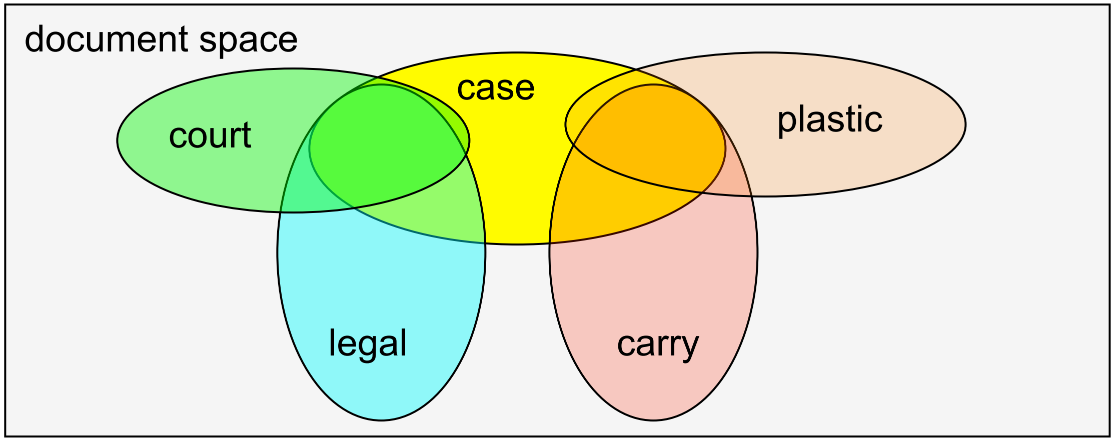
$10’000 Project

COMPLEX CASE STUDY:
CROWDSOURCING WORDNET

Biemann, C. (2012): Creating a system for lexical substitutions from scratch using crowdsourcing. Lang. Resources & Evaluation Vol. 47, No. 1, pp. 97--112. Springer. DOI 10.1007/s10579-012-9180-5

Word Senses and Keyword Search

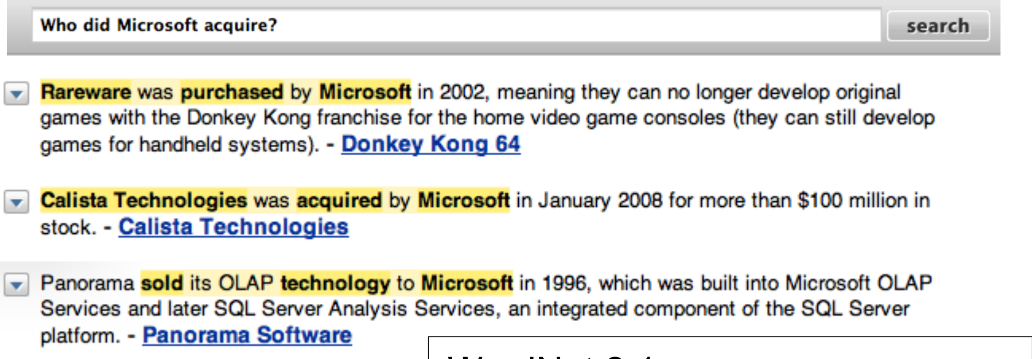
- Ambiguous words have multiple senses, e.g. *case* (container/legal/...). Mostly, sense frequency distribution is highly skewed in both collection and queries.
- Almost impossible to determine intended sense in short keyword queries
- "query word collocation effect" (Sanderson 2000): Combination of query words disambiguate each other



Sanderson, M. (2000): Retrieving with good sense. Information Retrieval, 2(1):49-69

Word Sense and Index expansion

- Semantic Search: Synonym matching through index expansion
- Spurious matches in absence of disambiguation
- Word sense handling enables semantic matching without spurious expansions, e.g. “foreign relations” - “external links”

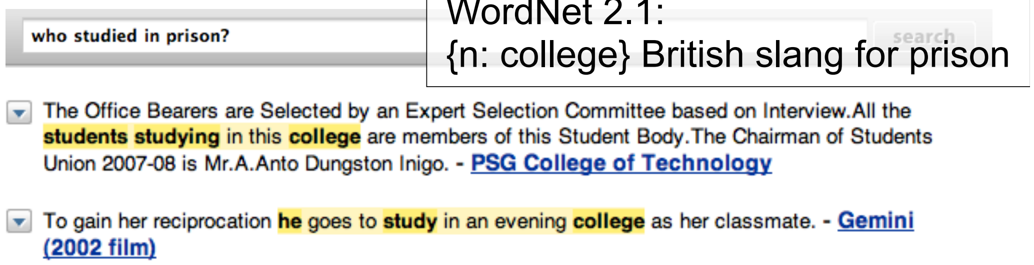


Who did Microsoft acquire?

Rareware was purchased by Microsoft in 2002, meaning they can no longer develop original games with the Donkey Kong franchise for the home video game consoles (they can still develop games for handheld systems). - Donkey Kong 64

Calista Technologies was acquired by Microsoft in January 2008 for more than $100 million in stock. - Calista Technologies

Panorama sold its OLAP technology to Microsoft in 1996, which was built into Microsoft OLAP Services and later SQL Server Analysis Services, an integrated component of the SQL Server platform. - Panorama Software

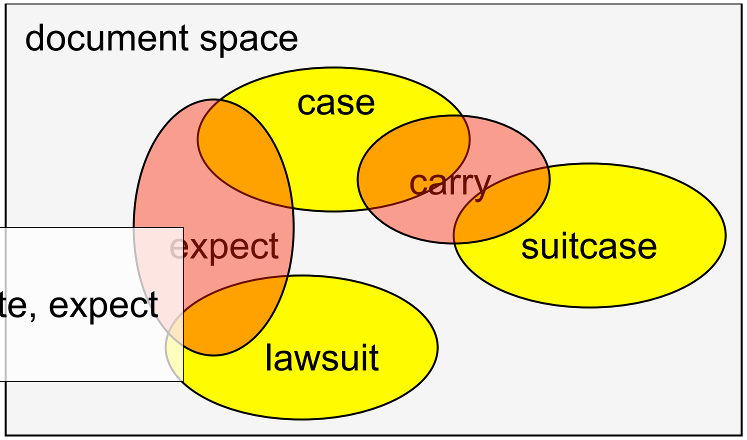


who studied in prison?

The Office Bearers are Selected by an Expert Selection Committee based on Interview.All the students studying in this college are members of this Student Body.The Chairman of Students Union 2007-08 is Mr.A.Anto Dungston Inigo. - PSG College of Technology

To gain her reciprocation he goes to study in an evening college as her classmate. - Gemini (2002 film)

WordNet 2.1:
{n: college} British slang for prison

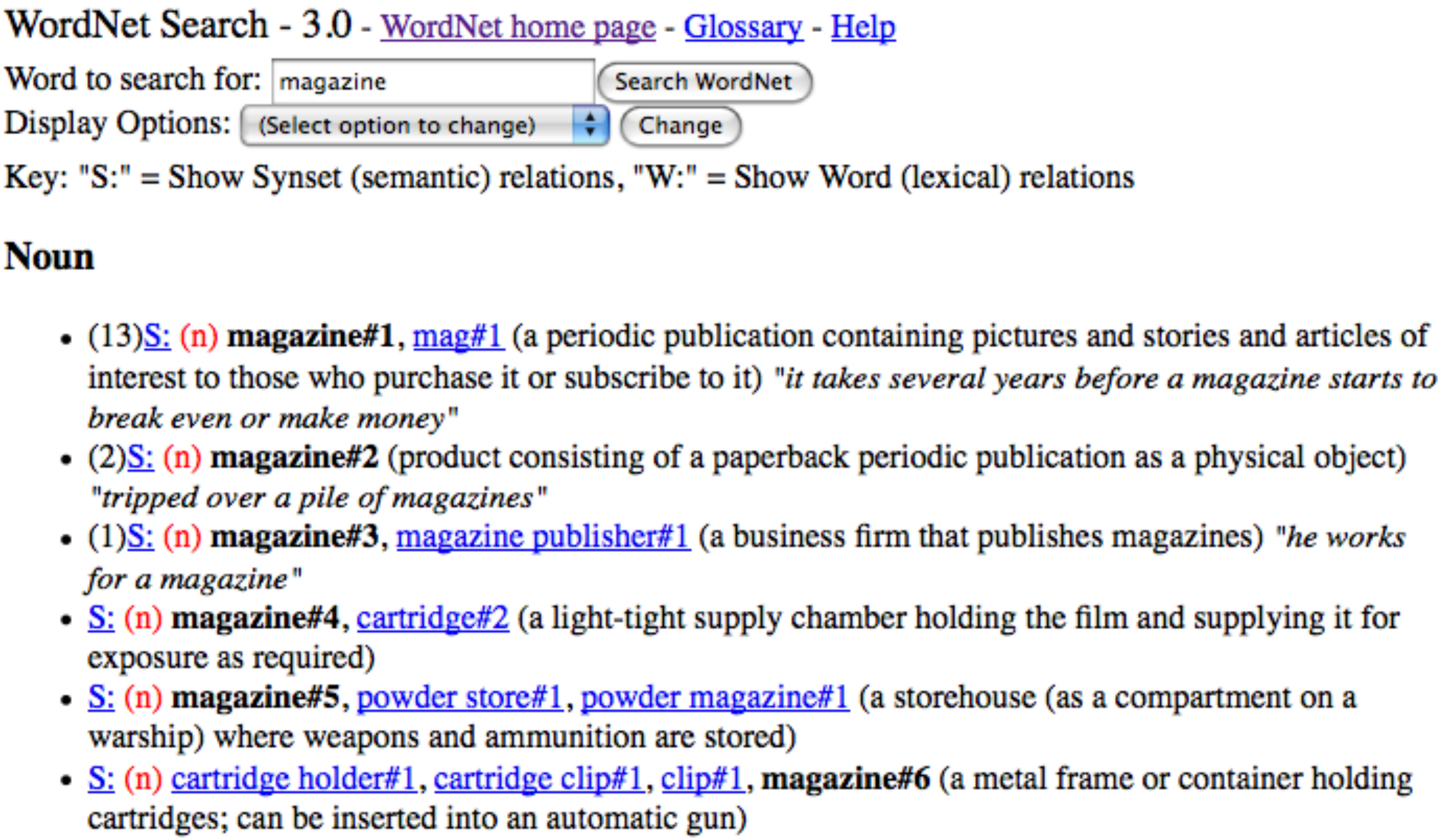


WordNet 3.0:
{v} bear, carry, gestate, expect
(be pregnant with)

Query: carry case

The Problem with WordNet

http://wordnetweb.princeton.edu/perl/webwn



WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: magazine

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- (13)S: (n) **magazine#1**, mag#1 (a periodic publication containing pictures and stories and articles of interest to those who purchase it or subscribe to it) *"it takes several years before a magazine starts to break even or make money"*
- (2)S: (n) **magazine#2** (product consisting of a paperback periodic publication as a physical object) *"tripped over a pile of magazines"*
- (1)S: (n) **magazine#3**, magazine publisher#1 (a business firm that publishes magazines) *"he works for a magazine"*
- S: (n) **magazine#4**, cartridge#2 (a light-tight supply chamber holding the film and supplying it for exposure as required)
- S: (n) **magazine#5**, powder store#1, powder magazine#1 (a storehouse (as a compartment on a warship) where weapons and ammunition are stored)
- S: (n) cartridge holder#1, cartridge clip#1, clip#1, **magazine#6** (a metal frame or container holding cartridges; can be inserted into an automatic gun)

- Systematic polysemy, obscure senses, little data to learn triggers from, granularity unclear

Building a sense labeled corpus

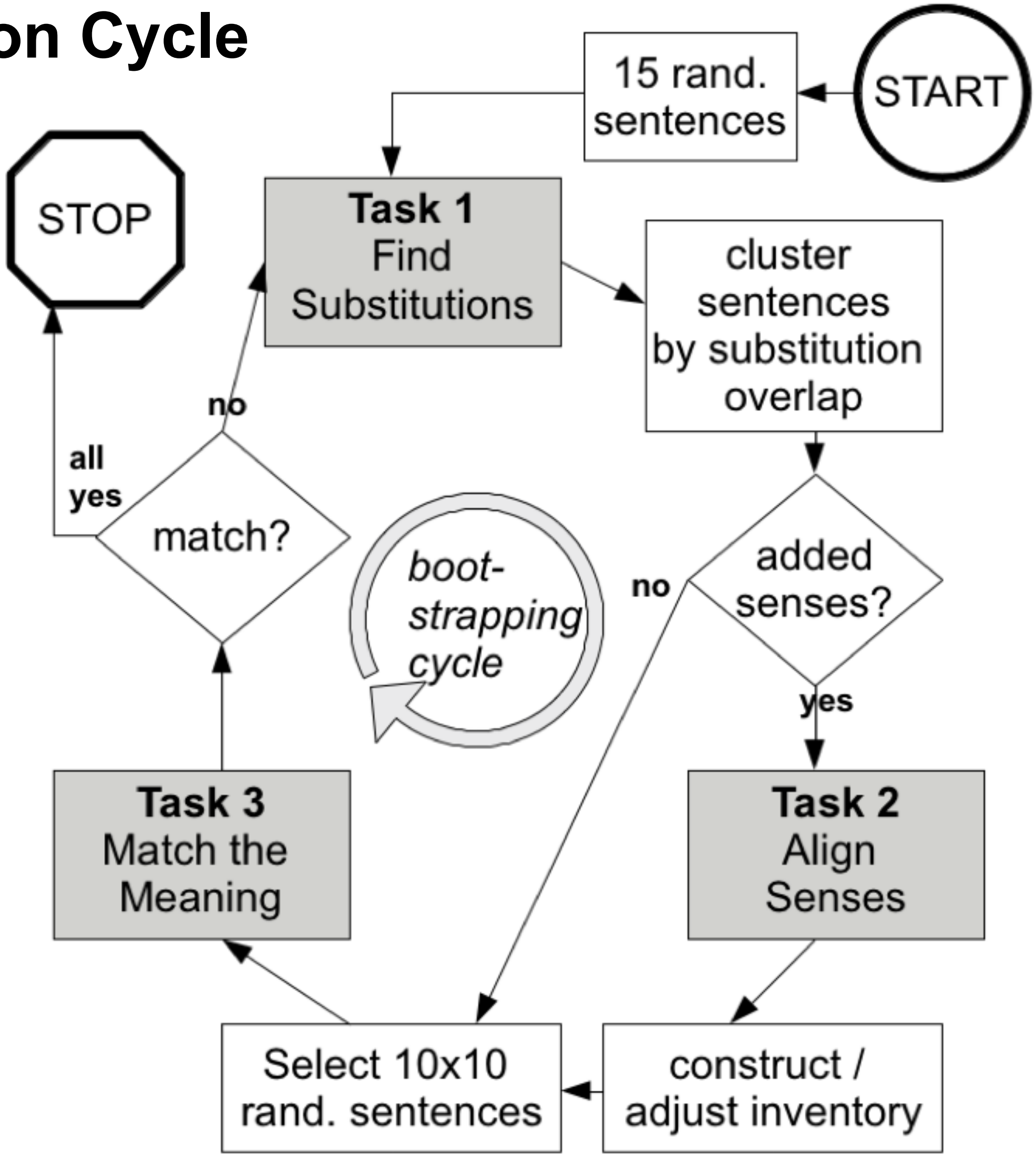
For a considerable amount of high frequency target words, we want

- Sentences containing these words, grouped by target word meaning
- Substitutes of target words, grouped by target word meaning

Desiderata:

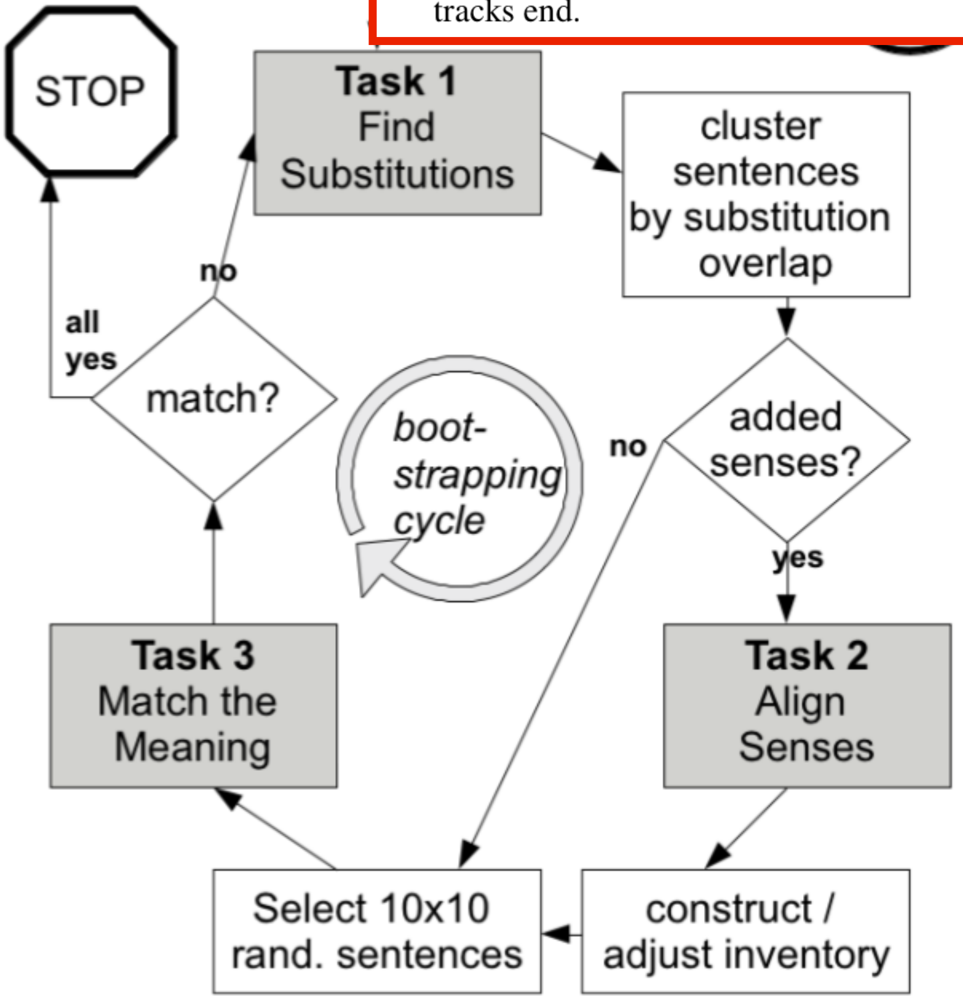
- Distribution (and coverage) of senses should match the underlying corpus
- Senses to be grouped by same substitutes rather than real world entities or such.

Acquisition Cycle



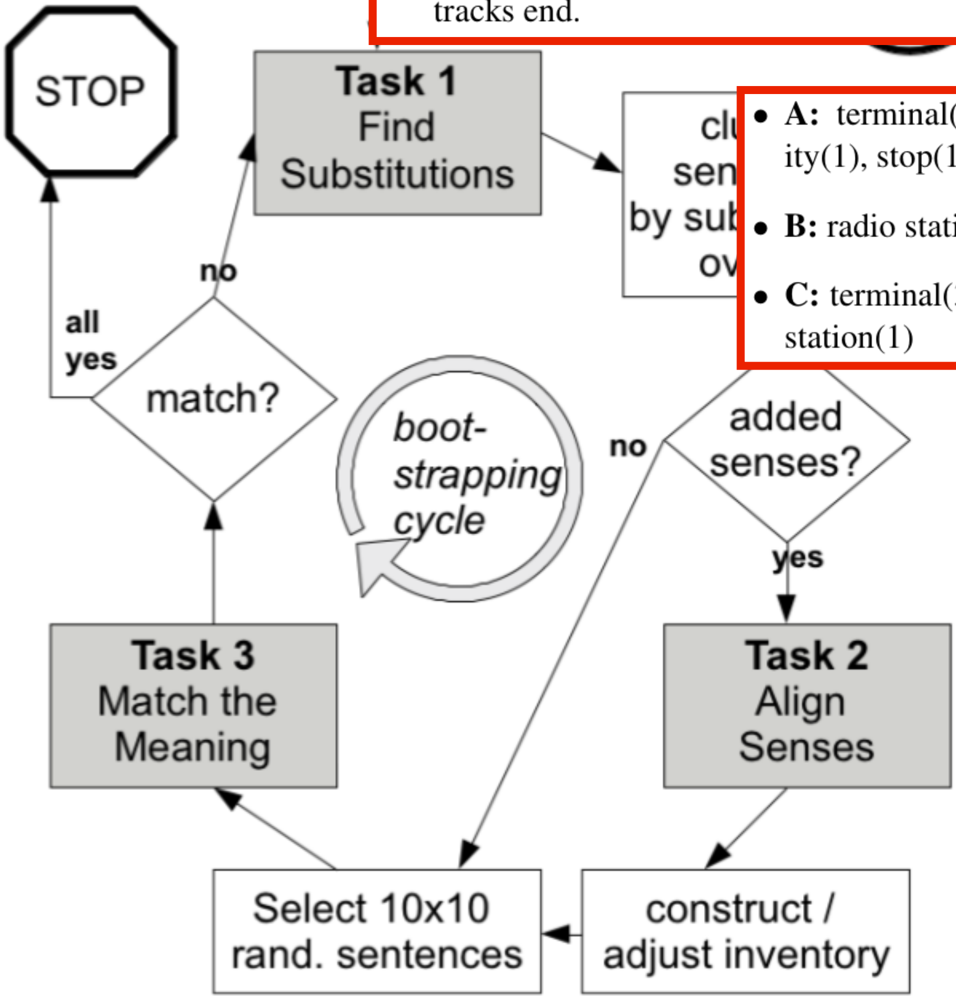
Sample run

- **A:** The train left the *station*.
- **B:** This radio *station* broadcasts news at five.
- **C:** Five miles from the *station*, the railway tracks end.



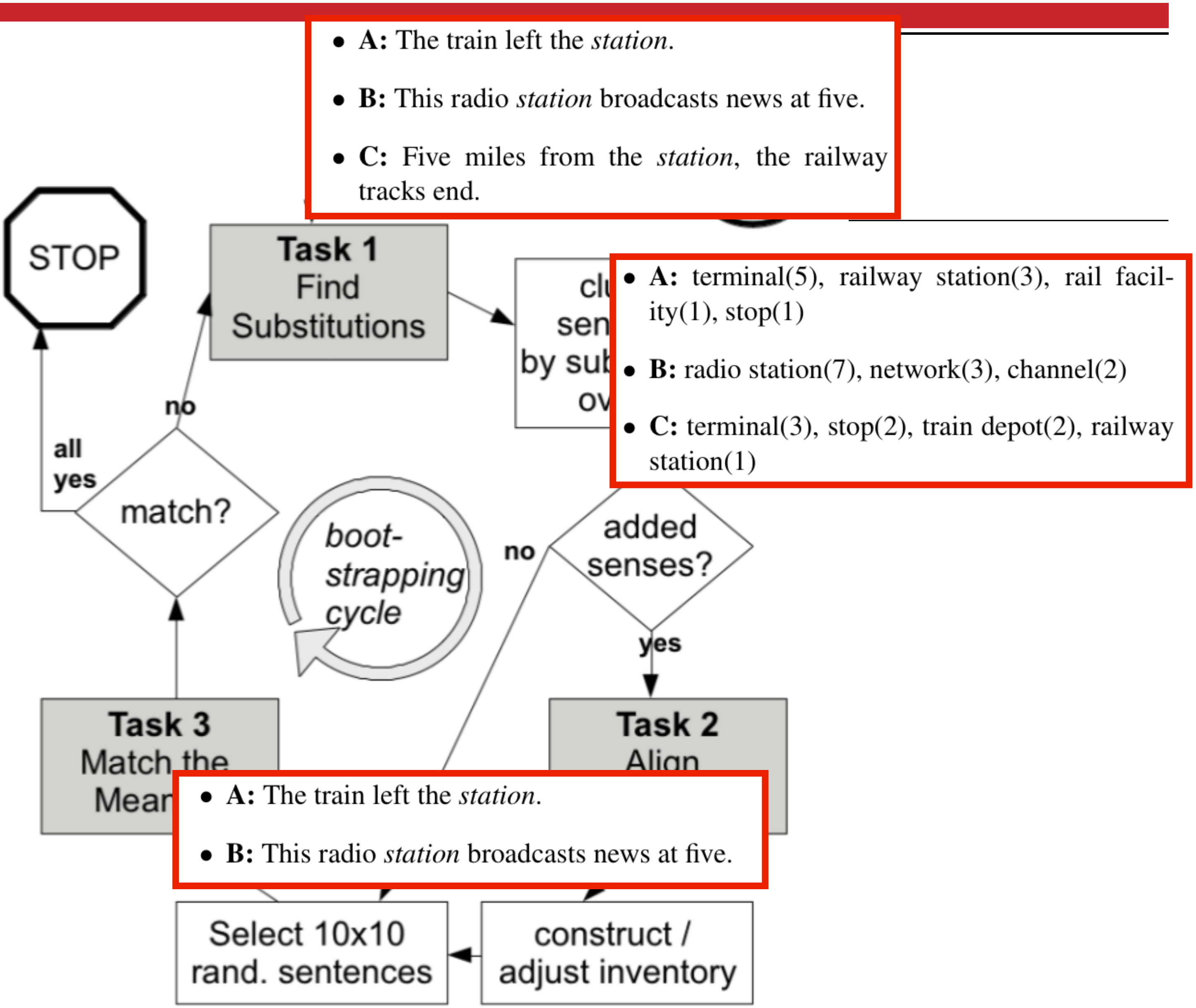
Sample run

- **A:** The train left the *station*.
- **B:** This radio *station* broadcasts news at five.
- **C:** Five miles from the *station*, the railway tracks end.



- **A:** terminal(5), railway station(3), rail facility(1), stop(1)
- **B:** radio station(7), network(3), channel(2)
- **C:** terminal(3), stop(2), train depot(2), railway station(1)

Sample run

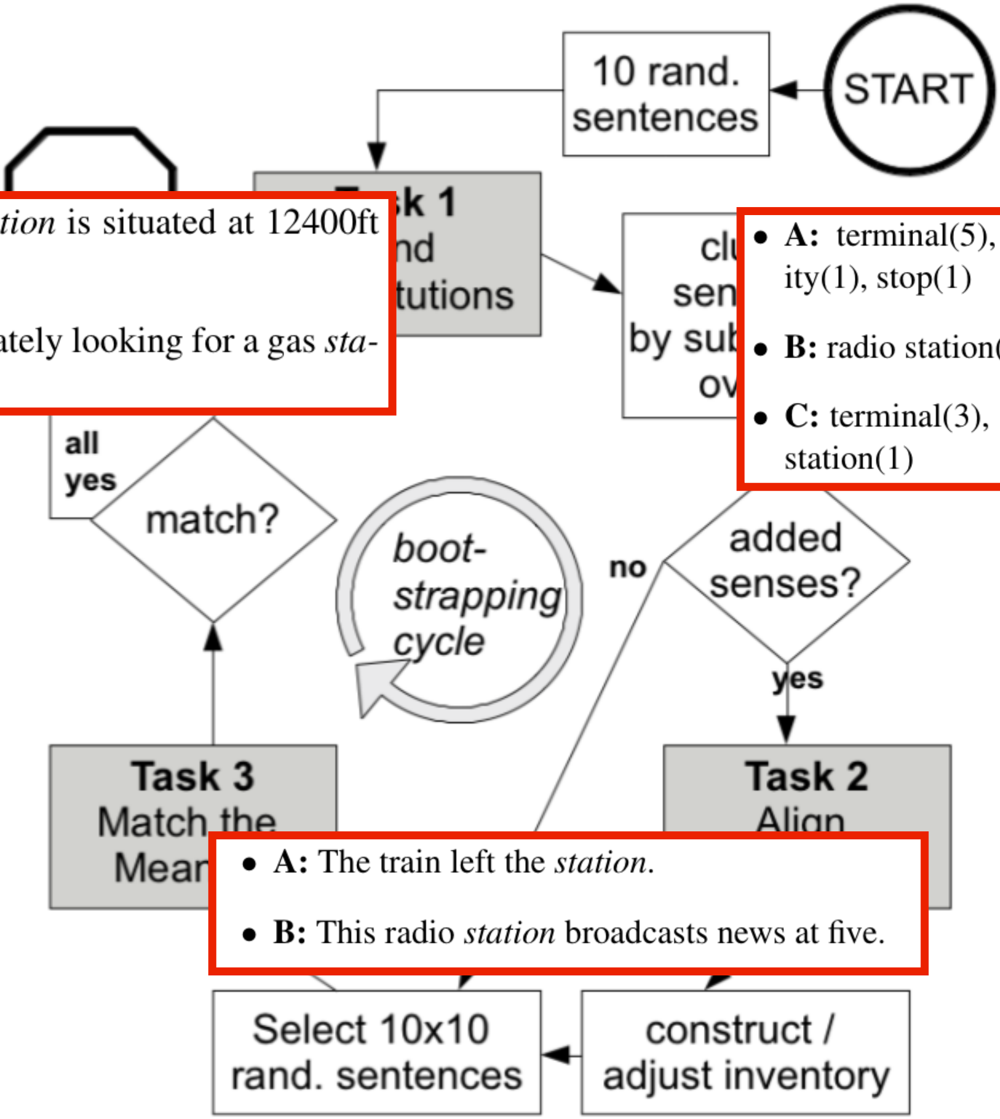


- **A:** The train left the *station*.
- **B:** This radio *station* broadcasts news at five.
- **C:** Five miles from the *station*, the railway tracks end.

- **A:** terminal(5), railway station(3), rail facility(1), stop(1)
- **B:** radio station(7), network(3), channel(2)
- **C:** terminal(3), stop(2), train depot(2), railway station(1)

- **A:** The train left the *station*.
- **B:** This radio *station* broadcasts news at five.

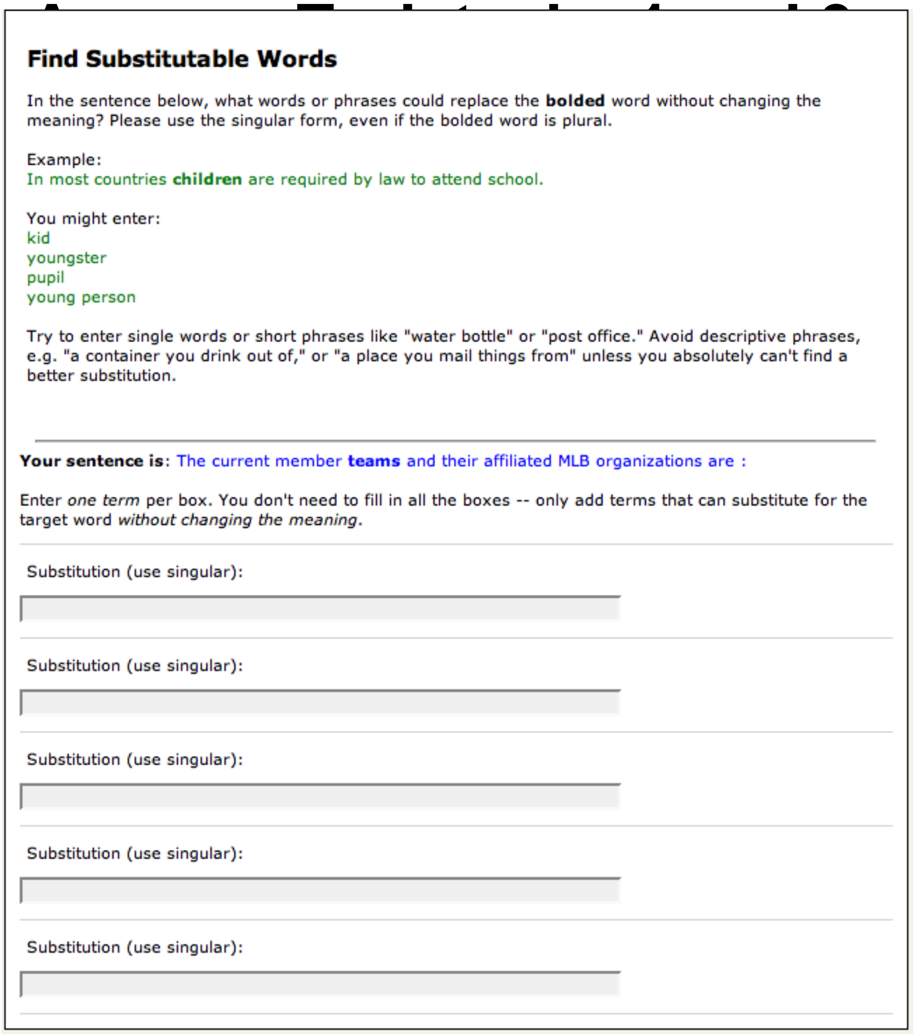
Sample run



- **D:** The mid-level *station* is situated at 12400ft altitude.
- **E:** They were desperately looking for a gas *station*.

- **A:** terminal(5), railway station(3), rail facility(1), stop(1)
- **B:** radio station(7), network(3), channel(2)
- **C:** terminal(3), stop(2), train depot(2), railway station(1)

- **A:** The train left the *station*.
- **B:** This radio *station* broadcasts news at five.



Find Substitutable Words

In the sentence below, what words or phrases could replace the **bolded** word without changing the meaning? Please use the singular form, even if the bolded word is plural.

Example:
In most countries **children** are required by law to attend school.

You might enter:
kid
youngster
pupil
young person

Try to enter single words or short phrases like "water bottle" or "post office." Avoid descriptive phrases, e.g. "a container you drink out of," or "a place you mail things from" unless you absolutely can't find a better substitution.

Your sentence is: The current member **teams** and their affiliated MLB organizations are :

Enter *one term* per box. You don't need to fill in all the boxes -- only add terms that can substitute for the target word *without changing the meaning*.

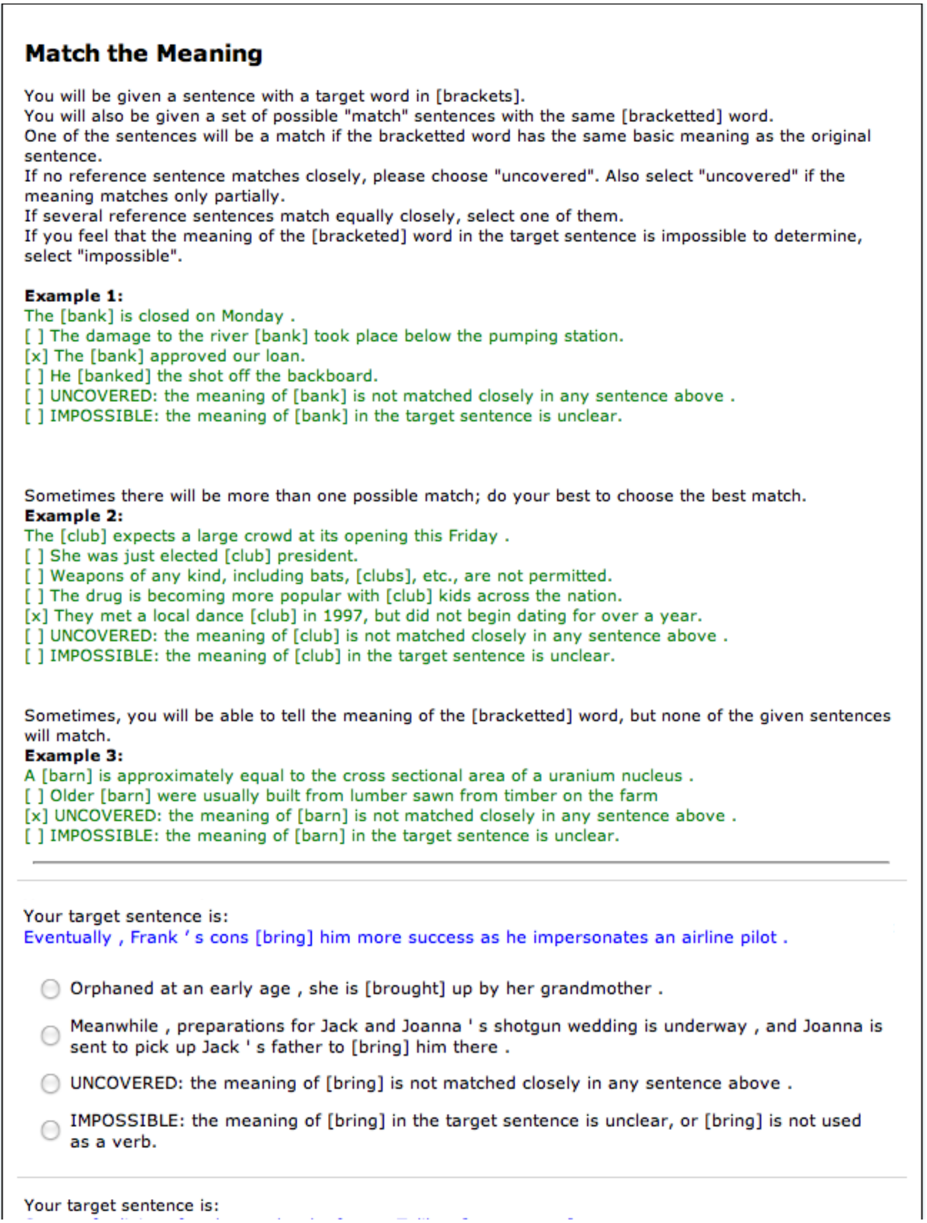
Substitution (use singular):

Substitution (use singular):

Substitution (use singular):

Substitution (use singular):

Substitution (use singular):



Match the Meaning

You will be given a sentence with a target word in [brackets].
You will also be given a set of possible "match" sentences with the same [bracketted] word.
One of the sentences will be a match if the bracketted word has the same basic meaning as the original sentence.
If no reference sentence matches closely, please choose "uncovered". Also select "uncovered" if the meaning matches only partially.
If several reference sentences match equally closely, select one of them.
If you feel that the meaning of the [bracketed] word in the target sentence is impossible to determine, select "impossible".

Example 1:
The [bank] is closed on Monday .
[ ] The damage to the river [bank] took place below the pumping station.
[x] The [bank] approved our loan.
[ ] He [banked] the shot off the backboard.
[ ] UNCOVERED: the meaning of [bank] is not matched closely in any sentence above .
[ ] IMPOSSIBLE: the meaning of [bank] in the target sentence is unclear.

Sometimes there will be more than one possible match; do your best to choose the best match.
Example 2:
The [club] expects a large crowd at its opening this Friday .
[ ] She was just elected [club] president.
[ ] Weapons of any kind, including bats, [clubs], etc., are not permitted.
[ ] The drug is becoming more popular with [club] kids across the nation.
[x] They met a local dance [club] in 1997, but did not begin dating for over a year.
[ ] UNCOVERED: the meaning of [club] is not matched closely in any sentence above .
[ ] IMPOSSIBLE: the meaning of [club] in the target sentence is unclear.

Sometimes, you will be able to tell the meaning of the [bracketted] word, but none of the given sentences will match.
Example 3:
A [barn] is approximately equal to the cross sectional area of a uranium nucleus .
[ ] Older [barn] were usually built from lumber sawn from timber on the farm
[x] UNCOVERED: the meaning of [barn] is not matched closely in any sentence above .
[ ] IMPOSSIBLE: the meaning of [barn] in the target sentence is unclear.

Your target sentence is:
Eventually , Frank ' s cons [bring] him more success as he impersonates an airline pilot .

○ Orphaned at an early age , she is [brought] up by her grandmother .

○ Meanwhile , preparations for Jack and Joanna ' s shotgun wedding is underway , and Joanna is sent to pick up Jack ' s father to [bring] him there .

○ UNCOVERED: the meaning of [bring] is not matched closely in any sentence above .

○ IMPOSSIBLE: the meaning of [bring] in the target sentence is unclear, or [bring] is not used as a verb.

Your target sentence is:

Current Resource in Numbers

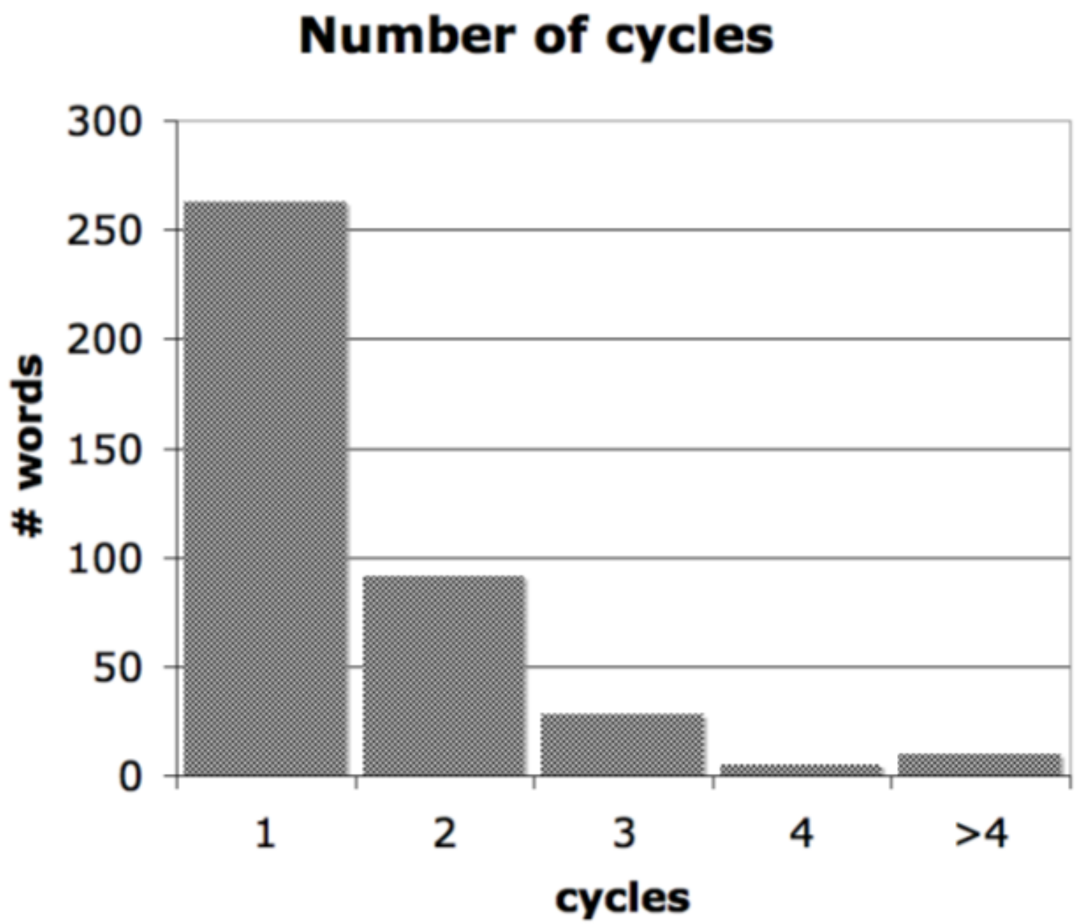
https://www.lt.informatik.tu-darmstadt.de/de/data/twsi-turk-bootstrap-word-sense-inventory/

- 1012 words (top frequent nouns in English Wikipedia), all but 50 from trusted turkers
- $8.30 cost per word on average
- 2.4 senses / word (WordNet: 6.3)
- avg. 100 sample sentences per sense
- 145’209 sentences with target word sense labels
- median: 10 substitutions per sense

During this project, a pool of ~50 trusted workers was built:

- used small, low-paying tasks and see how people perform
- granted qualifications to reliable and high-volume workers
- notified them on new patches per email
- paid bonuses for high volume and for rewriting the instructions

Cycles until convergence



Avg: 1.56 cycles per word

Quantitative Characteristics

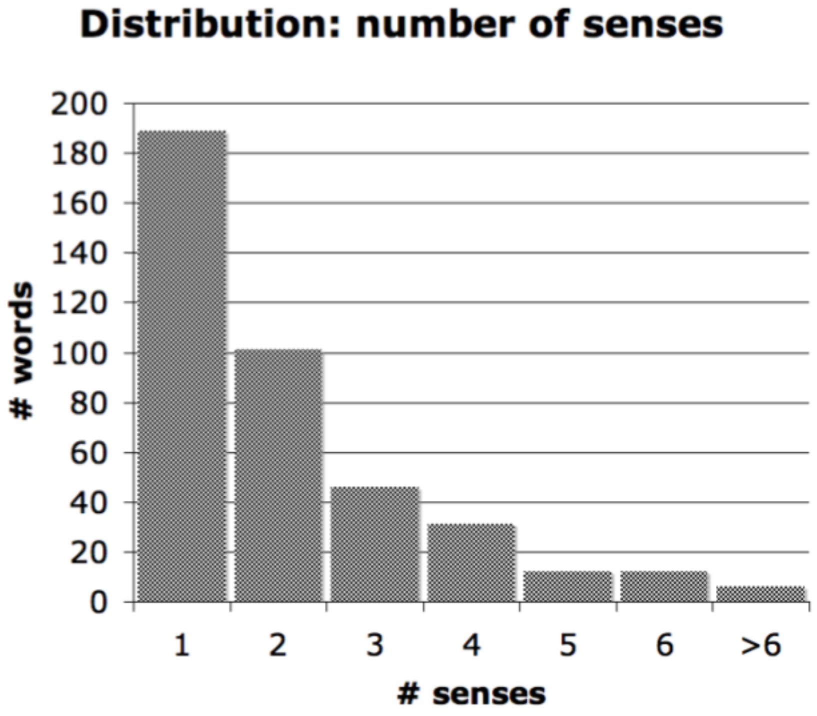


Figure 3: Distribution: number of words per number of senses. There are three words with 7 senses, two words with 9 senses and one word with 10 senses. The average is 2.1 senses per word.

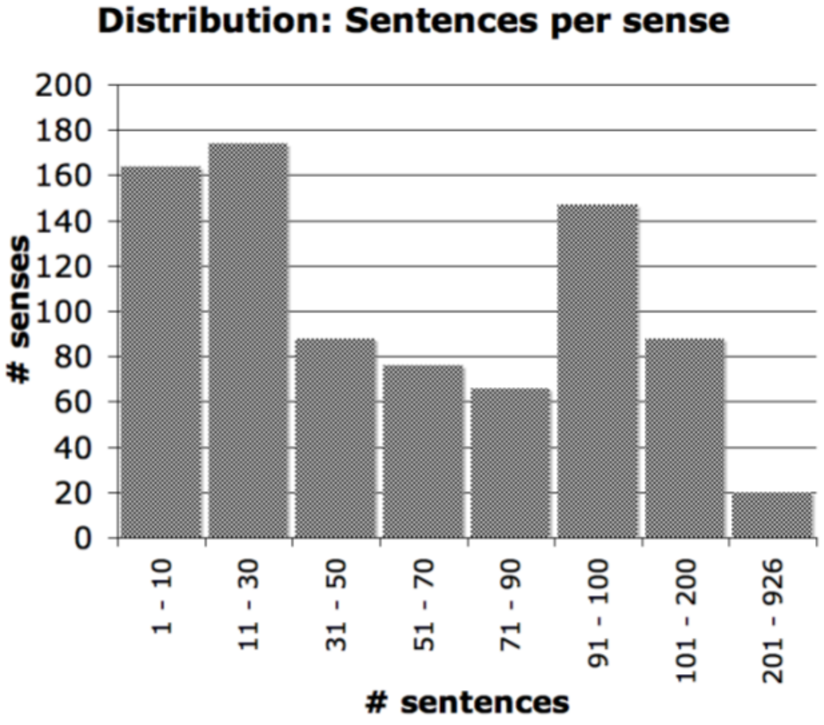


Figure 4: Distribution: number of senses per number of sentences interval. Only 5 senses have collected more than 300 sentences, at an average of 62.9 sentences per sense.

Example for reliable matches ‘case’

case@@1: The states and trade unions involved in the <b>case</b> have divided the issues between themselves , with the lawyers for each party arguing a particular part of the overall argument .

- The states and trade unions involved in the <b>case</b> have divided the issues ...
- Pressured by Rawls , Landsman orders Bunk to charge the <b>cases</b> .
- The Caledonian garnered attention in 2003 over a court <b>case</b> entitled Caledonian - ...
- ... Dutch tax police had visited Moscow in connection with the <b>case</b> .
- If the <b>case</b> is brought to the federal court under diversity jurisdiction ...
- Its payment is the second largest in the <b>case</b> , exceeded only by the $ 2.6 billion ...
- .. First Amendment student free speech <b>case</b> argued before the Supreme Court ...
- ... supporters began forming organizations to publicize the <b>case</b> and raise money for his defense .
- Frequently , civil <b>cases</b> are bifurcated into separate liability and damages proceedings .
- The Shah Bano <b>case</b> , where the Supreme Court recognised the Muslim woman ' s right ...
- ...

case@@2: For <b>cases</b> of fabricated evidence , the consequences can be wide ranging , with others working to confirm (or refute) the false finding , or with research agendas being distorted to address the fraudulent evidence

- Skeptics also point to historical <b>cases</b> in which flaws have been discovered ...
- ... , but Buffy does find herself with a " slight <b>case</b> of nudity. "
- In some <b>cases</b> , the larvae can develop on normal Drosophila lab ...
- ... helicopters may be used in <b>cases</b> where larger areas must be covered .
- ... , in the latter <b>case</b> only until an actual superior was elected or appointed .
- Such was the <b>case</b> on May 27 , 1780 , when a Swiss priest , ...
- In <b>case</b> a student is not content with the mark received , ...
- In this <b>case</b> , a door means a portal that connects one segment of a level to another ...
- ... , and as is the <b>case</b> generally in the world , the percentage of stamps ...
- ... ; in the <b>case</b> of adults , these are to be found on the underside of the thorax .
- ...

Error analysis (manual on 100 words)

- (4) Overlap or containment of one sense in the other leads to matching with two classes. An indicator of this is the number of set aside sentences in Task 3.
- (3) Systematic part-of speech tagger errors. E.g.: ”back”. Turkers did not consistently mark POS errors as impossible (although instructed). They reliably distinguished among senses.
- (3) Conflation of senses. Despite differences in meaning, two senses (as perceived by us) had sufficient overlap in their substitutions to not get clustered apart (e.g. ”relationship”)
- (2) Oscillation of senses. Differences in the subjective judgment of turkers caused the sense inventory to oscillate between grouping and distinguishing senses, such as ”over the centuries” vs. ”the 16th century”. Larger teams help.

Minor senses

- (8) Minor senses in set aside sentences (sampling).
  Some minor senses in the domain did not make it into the sense inventory; however, the cases never represented more than 4% of sample sentences.
- Few substitutions for minor senses. Of the 834 senses distinguished in our experiment, 41 did not get any substitution with frequency ≥ 2 and 142 senses did not record a substitution frequency of four or more.

For most applications, we are not interested in minor senses.

Magazine: the solution

- [189 sentences] magazine@@1

  Their first album was released by Columbia Records in 1972 , and they were voted " Best New Band " by Creem **magazine**.

  publication [42], periodical [32], journal [30], manual [9], gazette [5], newsletter [4], annual [3], digest [3], circular [2]

- [5 sentences] magazine@@2

  Instead , the film is pulled through the camera solely through the power of camera sprockets until the end , at which point springs or belts in the camera **magazine** pull the film back to the take - up side.

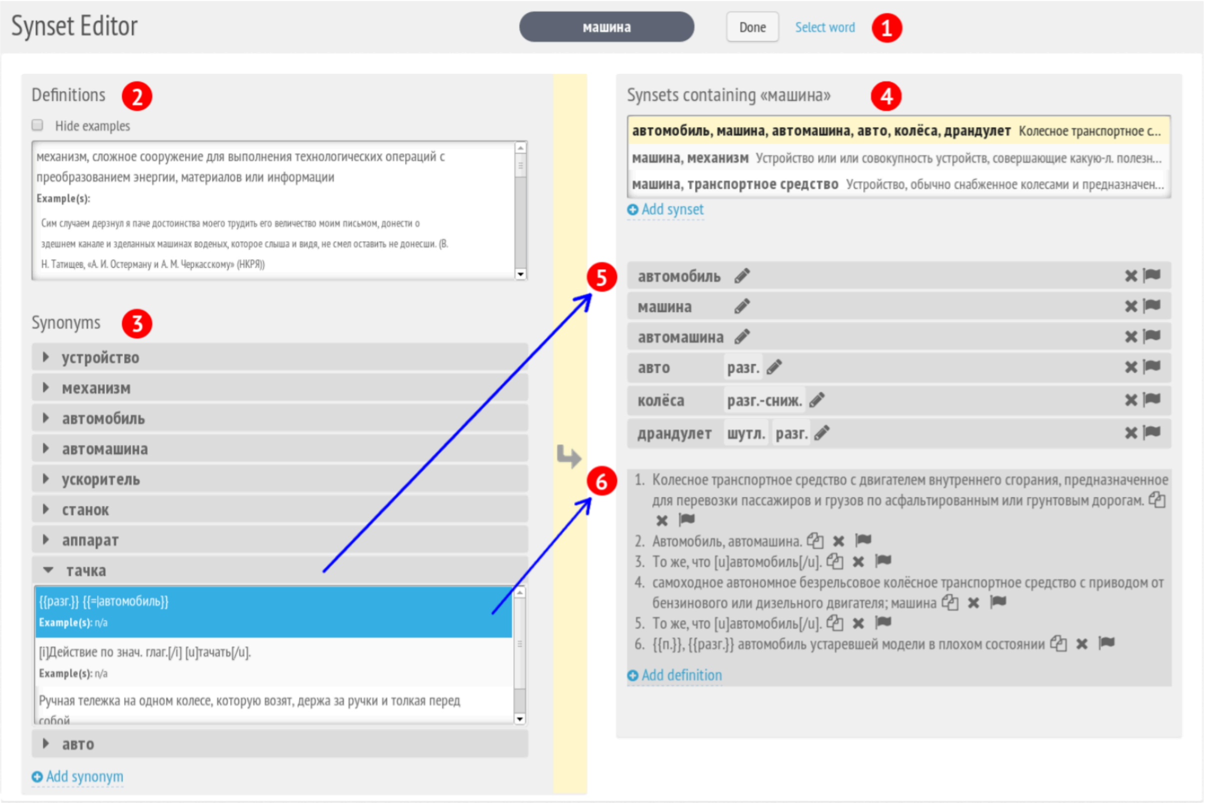
  cartridge [6], clip [5], chamber [3], holder [3], mag [3], ammunition chamber [2], cache [2], loading chamber [2]

Application: Sense-based Substitution System

https://www.lt.informatik.tu-darmstadt.de/de/software/twsi-sense-substituter/

Darmstadt is a <target= “**city**" lemma= “city" sense= “city" c= “1.0" substitutions= “[town, 89] [metropolis, 50] [municipality, 40] [metropolitan area, 17] [urban area, 14] [village, 14] [urban, 13] [community, 12] [megalopolis, 12] [township, 10]"> **in the Bundesland (federal** <target= “**state**" lemma= “state" sense= “state@@3" c= “0.6666667" substitutions= “[government, 7] [province, 2]"> **) of Hesse in Germany , located in the southern** <target= “**part**" lemma= “part" sense= “part@@1" c= “1.0" substitutions= “[portion, 21] [section, 21] [area, 17] [region, 15] [piece, 14] [component, 13] [segment, 11] [side, 8] [division, 6] [element, 4] [unit, 4]"> **of the Rhine Main** <target= “**Area**" lemma= “Area" sense= “area" c= “1.0" substitutions= “[region, 65] [zone, 24] [district, 22] [location, 21] [place, 19] [section, 17] [territory, 16] [field, 14] [part, 14] [vicinity, 14]"> **. The sandy** <target= “**soils**" lemma= “soil" sense= “soil@@1" x= “1.0" substitutions= “[earth, 26] [dirt, 23] [ground, 8] [loam, 6] [land, 3] [topsoil, 2]"> **in the Darmstadt** <target= “area" lemma= “**area**" sense= “area" x= “1.0" substitutions= “[region, 65] [zone, 24] [district, 22] [location, 21] [place, 19] [section, 17] [territory, 16] [field, 14] [part, 14] [vicinity, 14]"> **, ill-suited for agriculture in** <target= “**times**" lemma= “time" sense= “time@@1" x= “0.5" substitutions= “[instance, 99] [occasion, 95] [period, 82] [moment, 60] [era, 50] [age, 24] [event, 23] [point, 22] [occurrence, 17] [duration, 16]"> **before industrial fertilisation , [2] prevented any larger** <target= “**settlement**" lemma= “settlement" sense= “settlement@@1" x= “1.0" substitutions= “[colony, 21] [community, 19] [village, 12] [town, 8] [hamlet, 6] [establishment, 5] [habitation, 5]"> **from developing , until the** <target= “**city**" lemma= “city" sense= “city" x= “1.0" substitutions= “[town, 89] [metropolis, 50] [municipality, 40] [metropolitan area, 17] [urban area, 14] [village, 14] [urban, 13] [community, 12] [megalopolis, 12] [township, 10]"> **became the** <target= “**seat**" lemma= “seat" sense= “seat@@1" confidence= “1.0" substitutions= “[position, 21] [post, 19] [place, 10] [spot, 10] [elected post, 7] [station, 5] [rank, 3] [chair position, 2] [seat of government, 2]"> **of the Landgraves of Hessen.**

YARN: use student volunteers for complex tasks



механизм, сложное сооружение для выполнения технологических операций с преобразованием энергии, материалов или информации

Example(s):

Сим случаем дерзнул я паче достоинства моего трудить его величество моим письмом, донести о здешнем канале и зделанных машинах воденых, которое слыша и видя, не смел оставить не донесши. (В. Н. Татищев, «А. И. Остерману и А. М. Черкасскому» (НКРЯ))

- устройство
- механизм
- автомобиль
- автомашина
- ускоритель
- станок
- аппарат
- тачка
  - {{разг.}} {{=|автомобиль}} **Example(s):** n/a
  - [i]Действие по знач. глаг.[/i] [u]тачать[/u]. **Example(s):** n/a
  - Ручная тележка на одном колесе, которую возят, держа за ручки и толкая перед собой
- авто

автомобиль, машина, автомашина, авто, колёса, драндулет Колесное транспортное с...

машина, механизм Устройство или или совокупность устройств, совершающие какую-л. полезн...

машина, транспортное средство Устройство, обычно снабженное колесами и предназначен...

автомобиль		
машина		
автомашина		
авто	разг.	
колёса	разг.-сниж.	
драндулет	шутл.	разг.

1. Колесное транспортное средство с двигателем внутреннего сгорания, предназначенное для перевозки пассажиров и грузов по асфальтированным или грунтовым дорогам.
2. Автомобиль, автомашина.
3. То же, что [u]автомобиль[/u].
4. самоходное автономное безрельсовое колёсное транспортное средство с приводом от бензинового или дизельного двигателя; машина
5. То же, что [u]автомобиль[/u].
6. {{п.}}, {{разг.}} автомобиль устаревшей модели в плохом состоянии

Figure 1: Main window of YARN synset assembly interface (interface captions are translated for convenience of readers into English; originally all interface elements are in Russian): 1) initial word; 2) definitions and examples of the initial word; 3) possible synonyms of the initial word with definitions and examples; 4) a list of synsets containing the initial word (active synset is highlighted); 5) words constituting the current synset; 6) definitions of the current synset. The arrows show how the information items from the left-hand side form synsets in the right-hand side.

- Successfully crowdsourced “Russian WordNet” YARN
- Crowdworkers: 45 linguistics students, unpaid volunteers
- 1390 synsets; 970 of them contain more than a single word (253 contain 2 words, 228 — 3 words, 207 — 4, 282 — 5+).
- Editors spent about two minutes on building a ‘non-trivial’ synset on average

Braslavski, P., Ustalov, D., Mukhin, M. (2013): A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 101–104, Gothenburg, Sweden

In a Nutshell: Learned in Lesson 3

- Quality assurance mechanisms
- Find-Fix-Verify design pattern for open-ended tasks
- The Human Processing Unit (HPU) as a computation device
- Large HPU project on crowdsourcing WordNet and lexical substitution