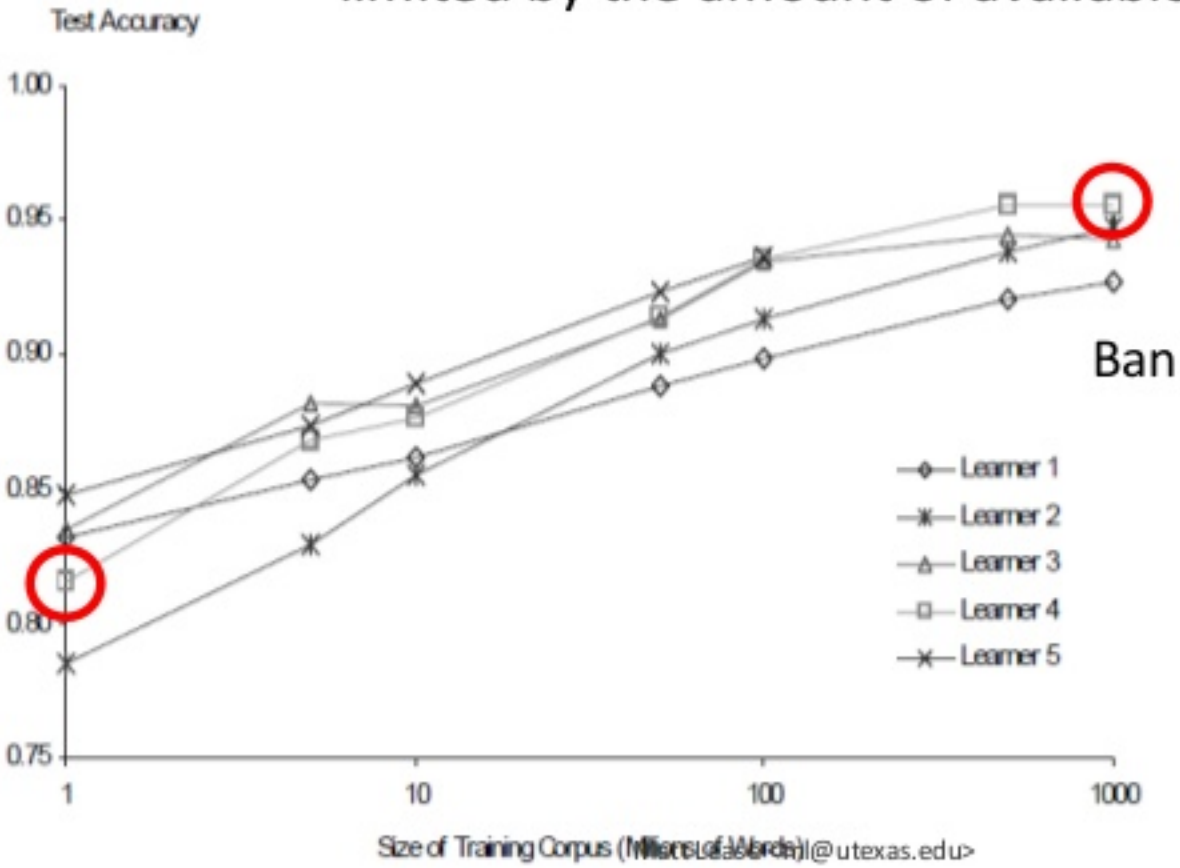
Data for NLP

The Unreasonable Effectiveness of Data

An AI system's effectiveness in practice is often limited by the amount of available training data



Banko and Brill (2001)

https://static.googleusercontent.com/media/research.google.com/uk//pubs/archive/35179.pdf

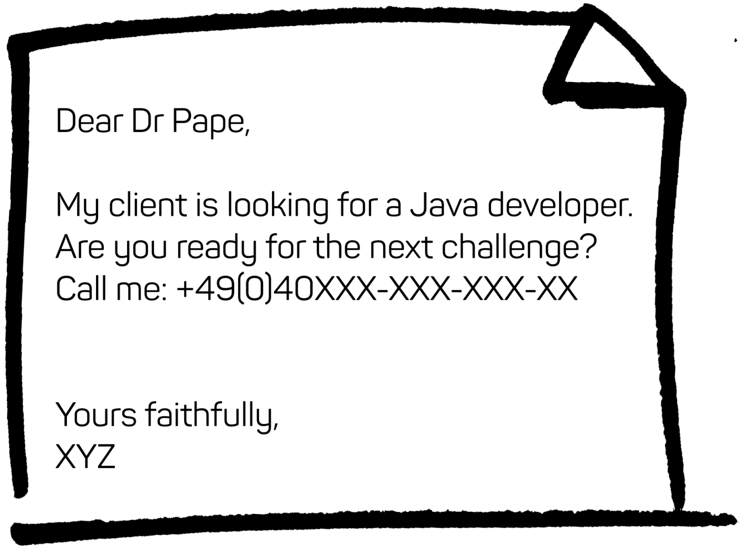
Uses for Language Data

Supervised learning – labeled data

Sentiment analysis

Spam detection

Intent analyzer



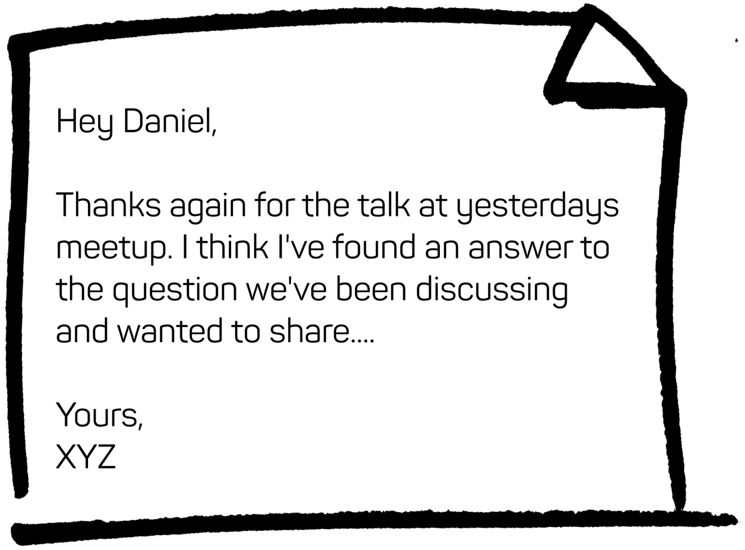
Dear Dr Pape,

My client is looking for a Java developer.
Are you ready for the next challenge?
Call me: +49(0)40XXX-XXX-XXX-XX

Yours faithfully,
XYZ

SPAM

vs.



Hey Daniel,

Thanks again for the talk at yesterdays meetup. I think I've found an answer to the question we've been discussing and wanted to share....

Yours,
XYZ

HAM

Uses for Language Data

For sequence-to-sequence tasks: a separate case of labeled language data where data units label each other

- machine translation
- NNs for chatbots

Data for NNs for chatbots:

“It’s my birthday today.”– “Happy birthday!”
“Happy birthday!” – “Thank you.”
“Thank you.” – “How old are you turning?”

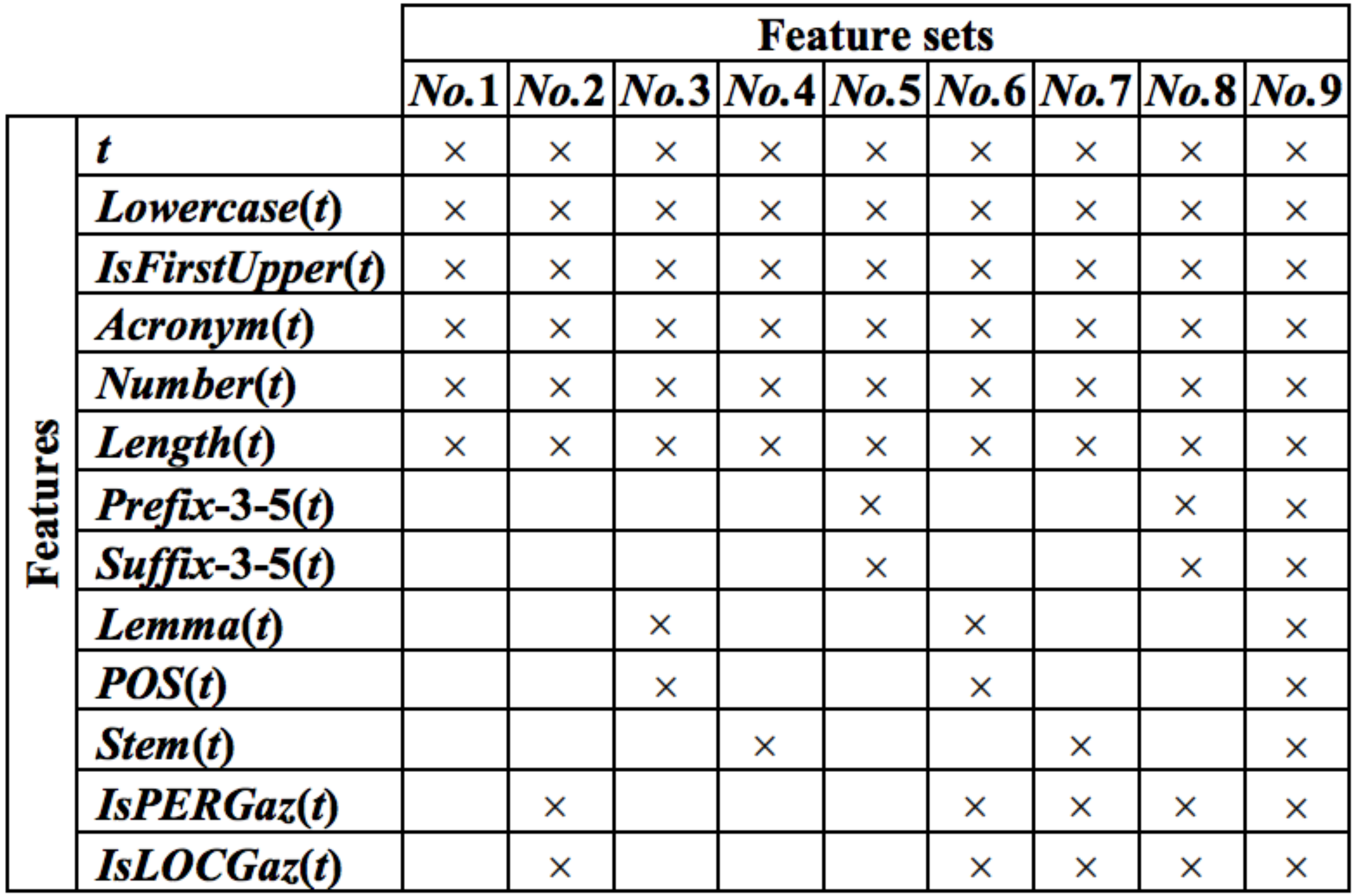
Uses for Language Data

Unlabeled, raw text data:

- Unsupervised learning (text clustering)
- Statistical language models
- NN language models (BERT, fasttext, word2vec)

Uses for Language Data

For linguistic analysis and feature engineering (i. e., manually inspect what phrases are frequent in a particular domain, what things seem to be informative, etc.).



		Feature sets								
		No.1	***No.2***	***No.3***	***No.4***	***No.5***	***No.6***	***No.7***	***No.8***	***No.9***
Features	***t***	×	×	×	×	×	×	×	×	×
	Lowercase(t)	×	×	×	×	×	×	×	×	×
	IsFirstUpper(t)	×	×	×	×	×	×	×	×	×
	Acronym(t)	×	×	×	×	×	×	×	×	×
	Number(t)	×	×	×	×	×	×	×	×	×
	Length(t)	×	×	×	×	×	×	×	×	×
	Prefix-3-5(t)					×			×	×
	Suffix-3-5(t)					×			×	×
	Lemma(t)			×			×			×
	POS(t)			×			×			×
	Stem(t)				×			×		×
	IsPERGaz(t)		×				×	×	×	×
	IsLOCGaz(t)		×				×	×	×	×

Uses for Language Data

Structured reference language data to expand the information about the training data. Used for feature engineering.

- annotated corpora
- gazetteers

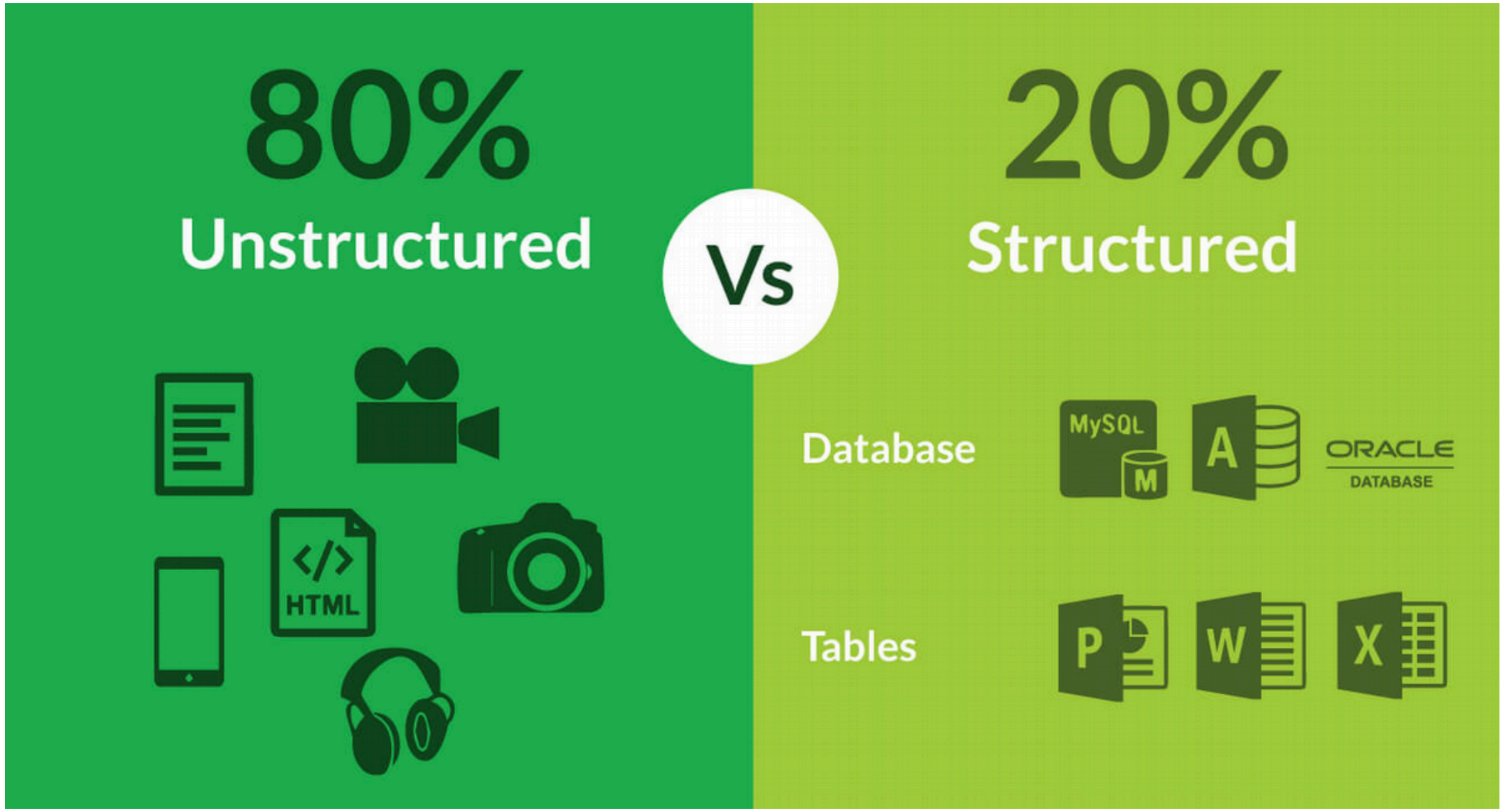
  Is this word a known proper name?
- ontologies and knowledge bases

  Is this word in an ontology fact like capitalOf(Kyiv, Ukraine)? If so, it's a city name.
- dictionaries, like WordNet, ConceptNet

  If we can't find the phrase "big expectations" in our statistical model, search the model with synonyms, like "great expectations".

Types of Data

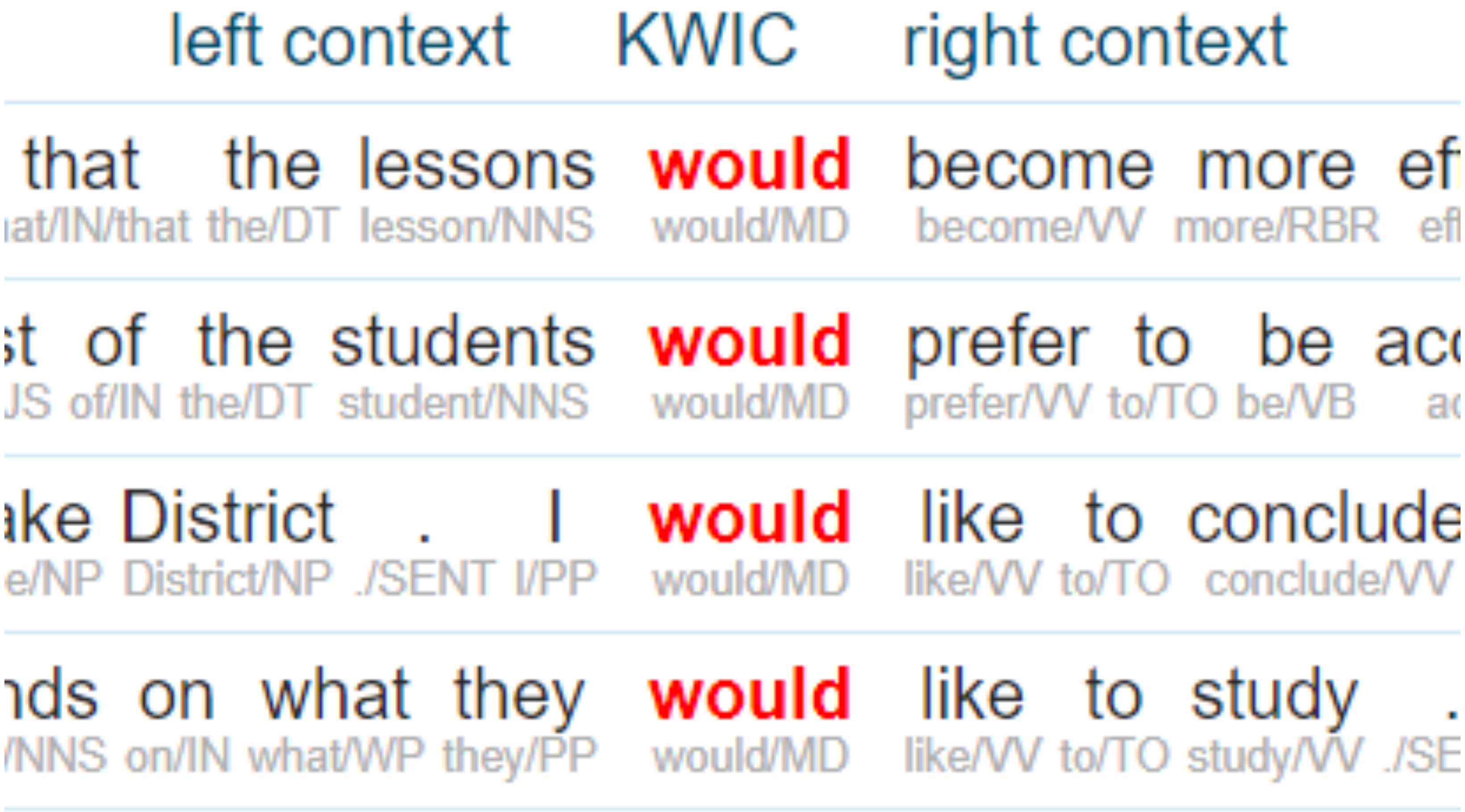
- Structured
- Semi-structured
- Unstructured



Structured Linguistic Data: Corpus

A corpus is an annotated collection of docs in a certain format.

Plural is **corpora**.



left context	KWIC	right context
that the lessons at/IN/that the/DT lesson/NNS	would would/MD	become more ef become/VV more/RBR ef
st of the students JS of/IN the/DT student/NNS	would would/MD	prefer to be acc prefer/VV to/TO be/VB ac
ake District . I e/NP District/NP ./SENT I/PP	would would/MD	like to conclude like/VV to/TO conclude/VV
nds on what they NNS on/IN what/WP they/PP	would would/MD	like to study . like/VV to/TO study/VV ./SE

Structured Linguistic Data: Corpus

Structured formats: Brown, BSF, PTB, XML, JSON, CSV

Brown format: word/POS-tag

```
    The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl said/vbd Friday/nr
an/at investigation/nn of/in Atlanta's/np$ recent/jj primary/nn election/nn
produced/vbd ``/`` no/at evidence/nn ''/'' that/cs any/dti irregularities/nns
took/vbd place/nn ./.


    The/at jury/nn further/rbr said/vbd in/in term-end/nn presentments/nns
that/cs the/at City/nn-tl Executive/jj-tl Committee/nn-tl ,/, which/wdt had/hvd
over-all/jj charge/nn of/in the/at election/nn ,/, ``/`` deserves/vbz the/at
praise/nn and/cc thanks/nns of/in the/at City/nn-tl of/in-tl Atlanta/np-tl ''/''
for/in the/at manner/nn in/in which/wdt the/at election/nn was/bedz
conducted/vbn ./.
```

Structured Linguistic Data: Corpus

SNLI corpus (JSONL+PTB):

Lisp-like dependency tree representations

```
{"annotator_labels": ["neutral", "entailment", "neutral",
"neutral", "neutral"], "captionID": "4705552913.jpg#2",
"gold_label": "neutral", "pairID": "4705552913.jpg#2r1n",
"sentence1": "Two women are embracing while holding to go
packages.", "sentence1_binary_parse": "( ( Two women )
( ( are ( embracing ( while ( holding ( to ( go packages ) )
) ) ) ) . ) )", "sentence1_parse": "(ROOT (S (NP (CD Two)
(NNS women)) (VP (VBP are) (VP (VBG embracing) (SBAR (IN
while) (S (NP (VBG holding)) (VP (TO to) (VP (VB go) (NP
(NNS packages)))))))) (. .)))", "sentence2": "The sisters
are hugging goodbye while holding to go packages after just
eating lunch.", "sentence2_binary_parse": "( ( The sisters )
( ( are ( ( hugging goodbye ) ( while ( holding ( to ( ( go
packages ) ( after ( just ( eating lunch ) ) ) ) ) ) ) ) ) .
) )", "sentence2_parse": "(ROOT (S (NP (DT The) (NNS
sisters)) (VP (VBP are) (VP (VBG hugging) (NP (UH goodbye))
(PP (IN while) (S (VP (VBG holding) (S (VP (TO to) (VP (VB
go) (NP (NNS packages)) (PP (IN after) (S (ADVP (RB just))
(VP (VBG eating) (NP (NN lunch))))))))))))) (. .)))"}
```

Useful Corpora Info

- National: OANC/MASC, British (non-free)
- LDC (non-free): Penn Treebank, OntoNotes, Web Treebank
- Books: Gutenberg, GoogleBooks
- Corporate: Reuters, Enron
- Research: SNLI, SquAD
- Multilang: UDeps, Europarl, European Commission Corpus (free): https://ec.europa.eu/jrc/en/language-technologies/dcep

Structured Linguistic Data: Ukrainian

Data for Ukrainian Language: lang-uk group

http://lang.org.ua/en/corpora/

NER corpus: https://github.com/lang-uk/ner-uk
Tonal dictionary: https://github.com/lang-uk/tonal-model
Gazetteers: https://github.com/lang-uk/ua-gazetteers

Corpora Cons

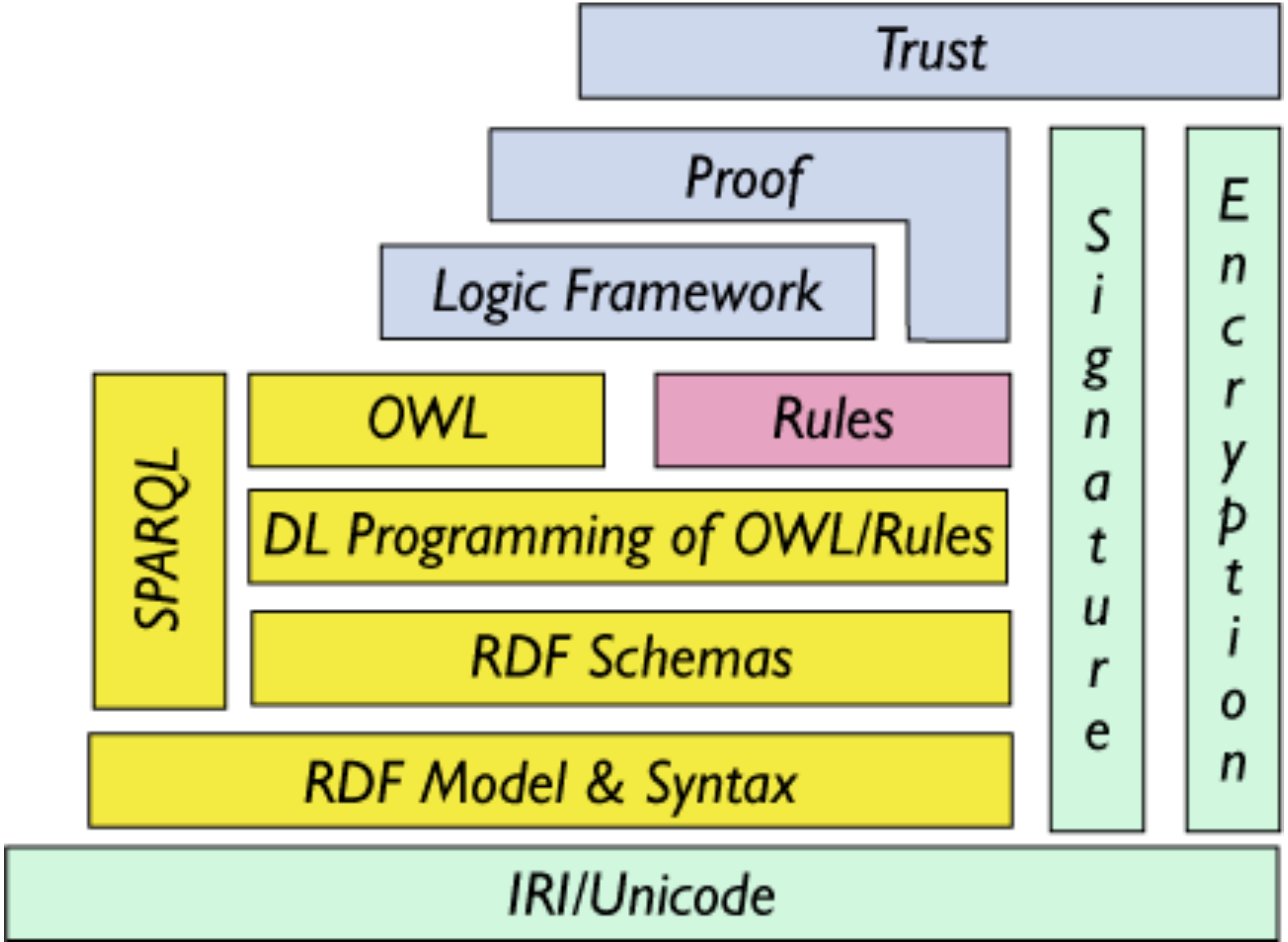
- Good corpora are not free and need licensing
- Contain language from a specific domain
- Annotation and structure usually contain errors
- Processing of custom formats is time-consuming

Structured Linguistic Data: DBs and KBs

Semantic Web:

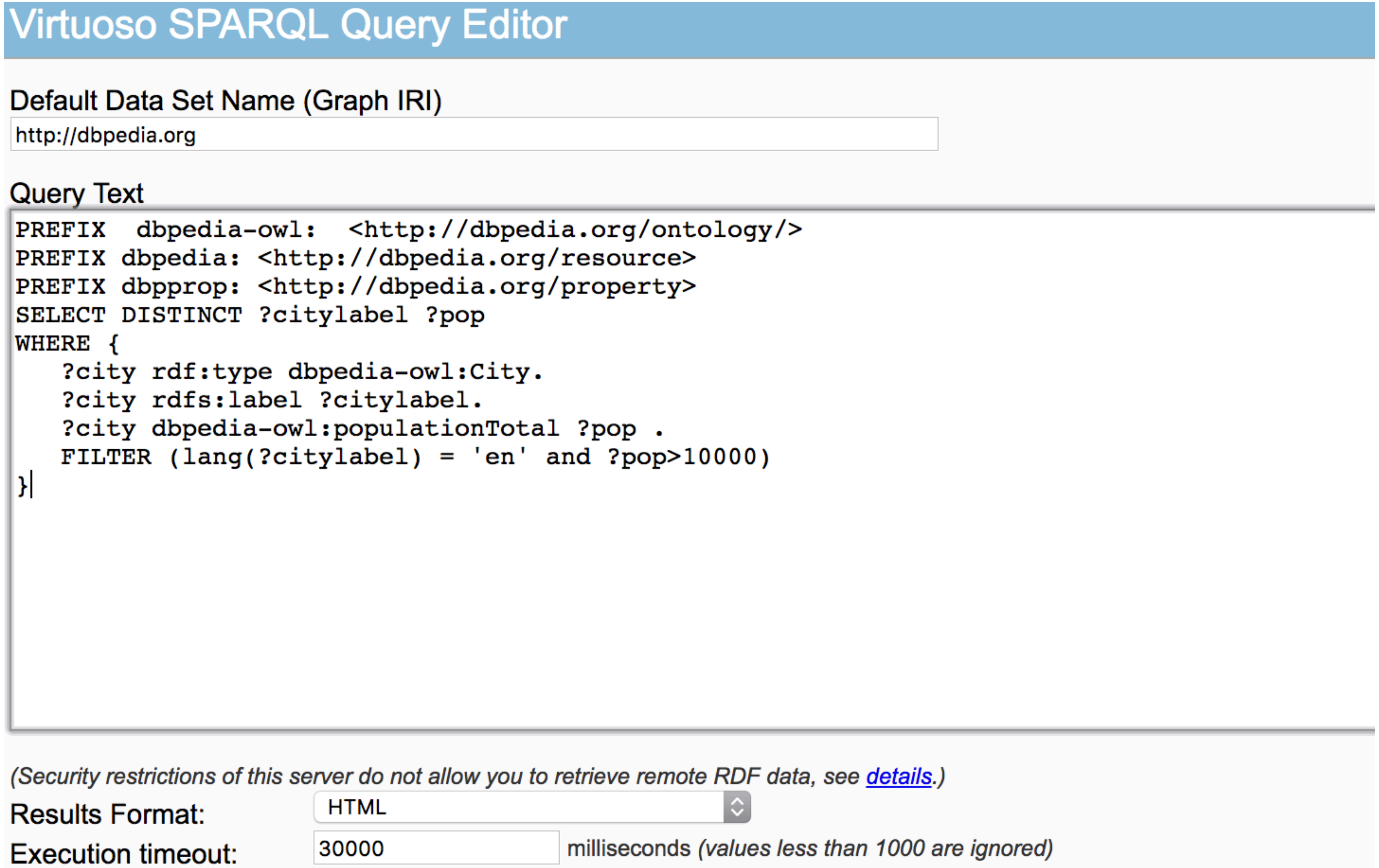
An effort to structure and easily share information from the internet.

RDF
RDFS
Rule Interchange Format (RIF)
SPARQL
Web Ontology Language (OWL)
XML



Structured Linguistic Data: DBs and KBs

Using SPARKQL to query DBpedia (structured Wikipedia data)



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://dbpedia.org

Query Text

```
PREFIX  dbpedia-owl:  <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource>
PREFIX dbpprop: <http://dbpedia.org/property>
SELECT DISTINCT ?citylabel ?pop
WHERE {
   ?city rdf:type dbpedia-owl:City.
   ?city rdfs:label ?citylabel.
   ?city dbpedia-owl:populationTotal ?pop .
   FILTER (lang(?citylabel) = 'en' and ?pop>10000)
}
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML

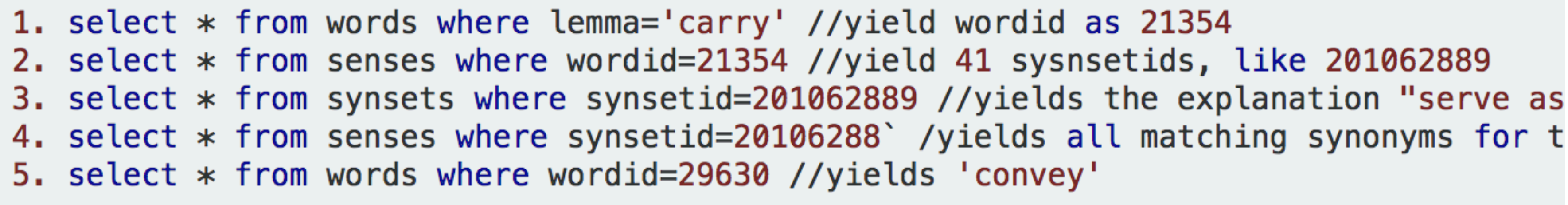
Execution timeout: 30000 milliseconds *(values less than 1000 are ignored)*

Structured Linguistic Data: Dictionaries

WordNet - a large lexical database of English. Contains synonyms, antonyms, semantic relations (hyponym - hyperonym).

How to access:

- NLTK WordNet interface
- DB queries



```
select * from words where lemma='carry' //yield wordid as 21354
select * from senses where wordid=21354 //yield 41 sysnsetids, like 201062889
select * from synsets where synsetid=201062889 //yields the explanation "serve as
select * from senses where synsetid=20106288` /yields all matching synonyms for t
select * from words where wordid=29630 //yields 'convey'
```

Unstructured Linguistic Data: Raw Text from Internet

Already scraped web-pages:

CluWeb: https://www.lemurproject.org/clueweb12.php/

Common Crawl: http://commoncrawl.org/

Raw text is easy to get but…

- Huge processing effort
- Large amount of errors
- Web noise

Problems With Available Data

Good data belongs to somebody and needs to be licensed.

Data owners:

Universities
Companies
Individuals

- Either low quality or expensive
- Nobody wants to share
- Legal reasons

I'LL MAKE MY OWN DATA
imgflip.com

How to Create Linguistic Data

- Scraping
- Annotation tools
- Crowdsourcing
- Generating yourself

Scraping

- Web-page scraping

- Extracting from non-HTML Formats (.pdf, .doc...)

- Getting from API
  - Twitter: pull tweets in real time (needs A LOT of preprocessing)
  - Webhose: scraped web-pages grouped into domains

Create Your Own Corpus: Corpus Annotation

Steps:

- Collect good-quality data to be annotated
- What is the end format of the corpus and annotation guidelines
- Pick the annotation tool
- Get people to annotate
- Analyze the quality
- Iterate

Who will annotate?

- Professional linguists (Appen)

  expensive

  good quality work
- Annotation monkeys (mturk)

  less expensive but not free

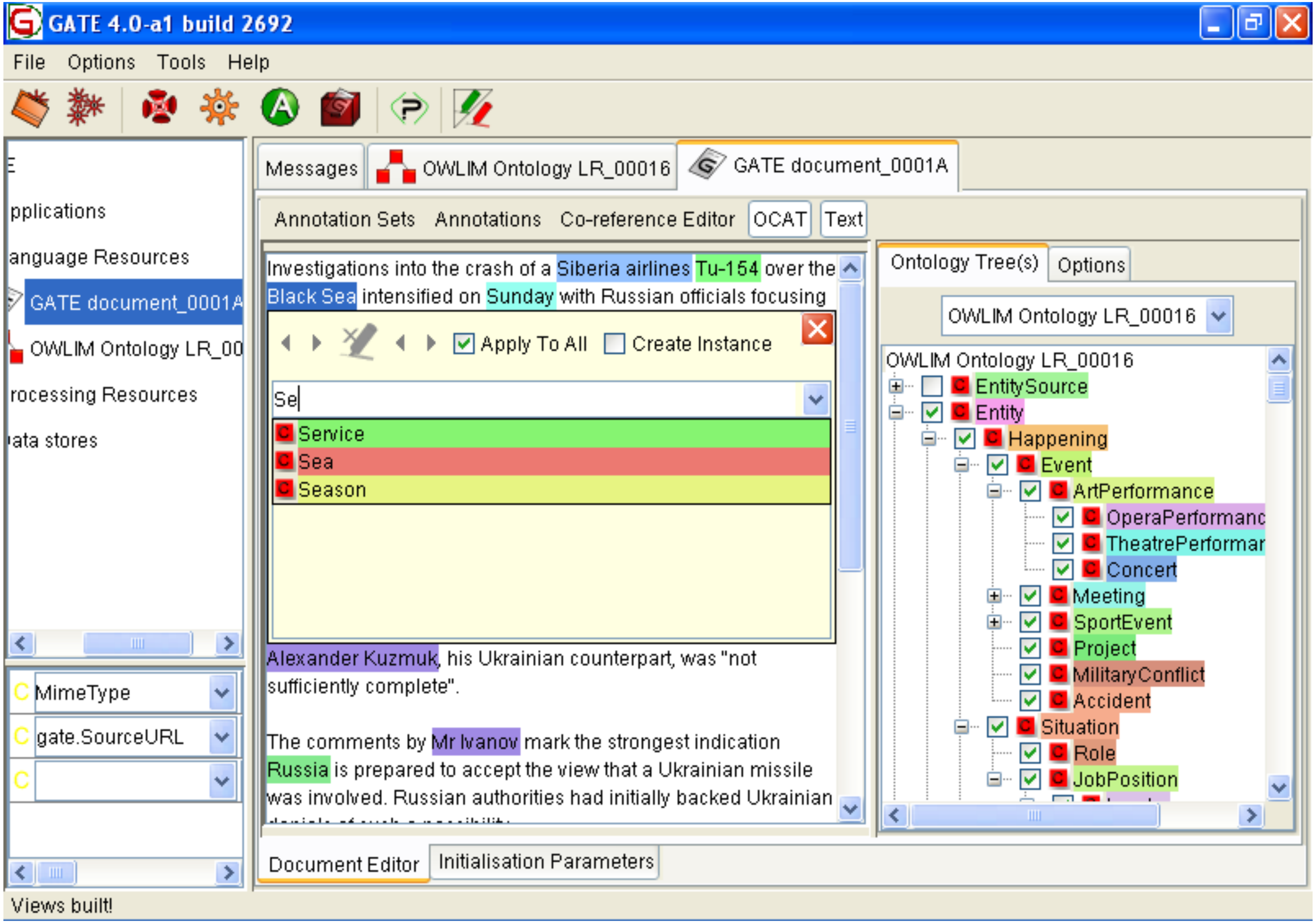
  prone to errors
- Volunteers (crowdsourcing)

  pretty much impossible

Annotation Tools

- Doccano
- Brat
- Anaphora
- Prodigy
- Anagram
- Vulyk (based on Brat)
- Ann
- GATE

GATE



Investigations into the crash of a Siberia airlines Tu-154 over the Black Sea intensified on Sunday with Russian officials focusing

☑ Apply To All □ Create Instance

Se

- Service
- Sea
- Season

Alexander Kuzmuk, his Ukrainian counterpart, was "not sufficiently complete".

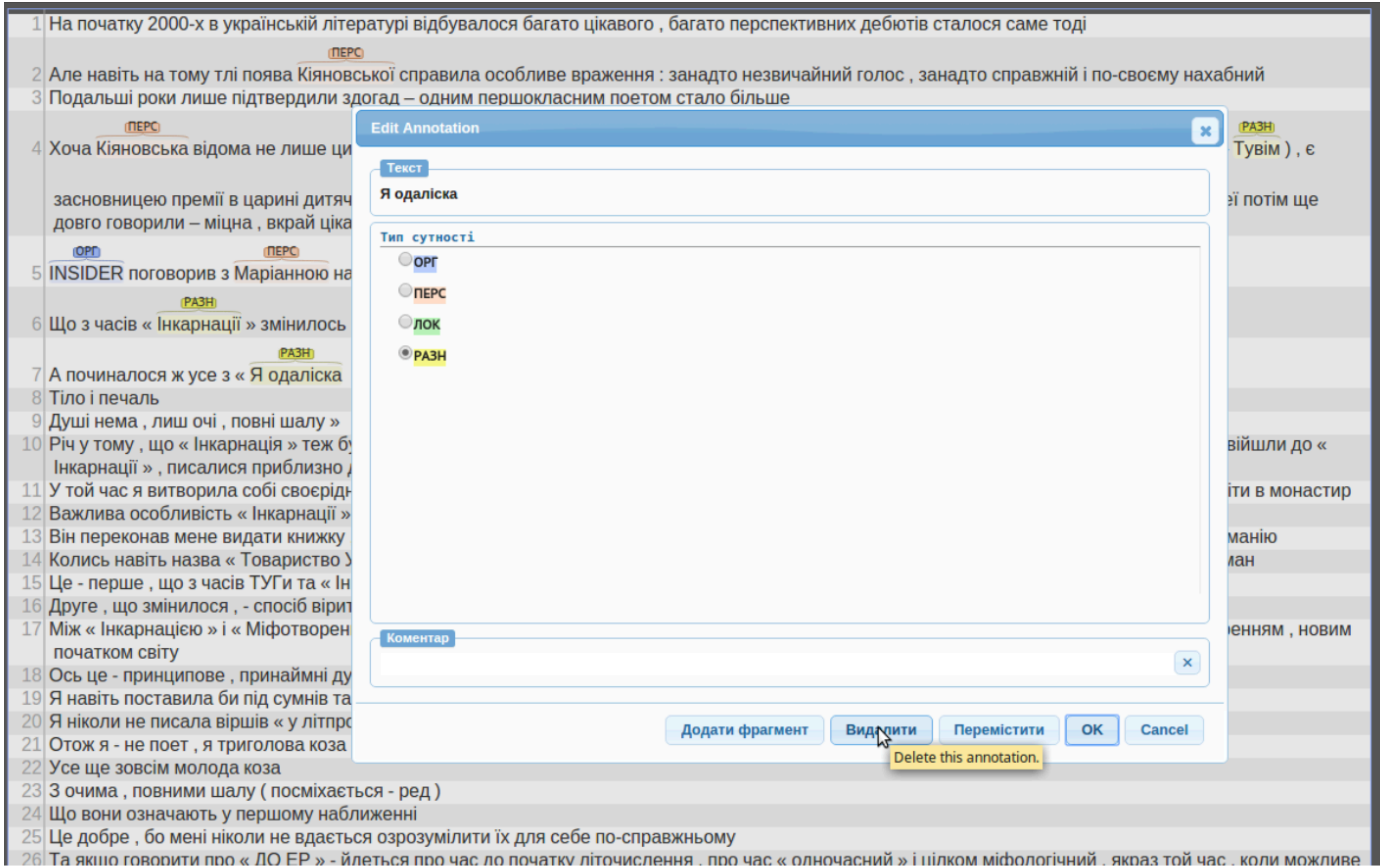
The comments by Mr Ivanov mark the strongest indication Russia is prepared to accept the view that a Ukrainian missile was involved. Russian authorities had initially backed Ukrainian denials of such a possibility.

C MimeType

C gate.SourceURL

Views built!

Vulyk: Ukrainian Free Annotation Tool



1 На початку 2000-х в українській літературі відбувалося багато цікавого , багато перспективних дебютів сталося саме тоді

2 Але навіть на тому тлі поява Кіяновської справила особливе враження : занадто незвичайний голос , занадто справжній і по-своєму нахабний

3 Подальші роки лише підтвердили здогад – одним першокласним поетом стало більше

4 Хоча Кіяновська відома не лише ци … Тувім) , є засновницею премії в царині дитяч … ї потім ще довго говорили – міцна , вкрай ціка

5 INSIDER поговорив з Маріанною на

6 Що з часів « Інкарнації » змінилось

7 А починалося ж усе з « Я одаліска

8 Тіло і печаль

9 Душі нема , лиш очі , повні шалу »

10 Річ у тому , що « Інкарнація » теж б … війшли до « Інкарнації » , писалися приблизно

11 У той час я витворила собі своєрідн … іти в монастир

12 Важлива особливість « Інкарнації »

13 Він переконав мене видати книжку … манію

14 Колись навіть назва « Товариство … ман

15 Це - перше , що з часів ТУГи та « Ін

16 Друге , що змінилося , - спосіб вірш

17 Між « Інкарнацією » і « Міфотворен … енням , новим початком світу

18 Ось це - принципове , принаймні ду

19 Я навіть поставила би під сумнів та

20 Я ніколи не писала віршів « у літпро

21 Отож я - не поет , я триголова коза

22 Усе ще зовсім молода коза

23 З очима , повними шалу (посміхається - ред)

24 Що вони означають у першому наближенні

25 Це добре , бо мені ніколи не вдається озрозуміліти їх для себе по-справжньому

26 Та якщо говорити про « ДО ЕР » - йдеться про час до початку літочислення , про час « одночасний » і цілком міфологічний , якраз той час , коли можливе

Edit Annotation ✕

Текст

Я одаліска

Тип сутності

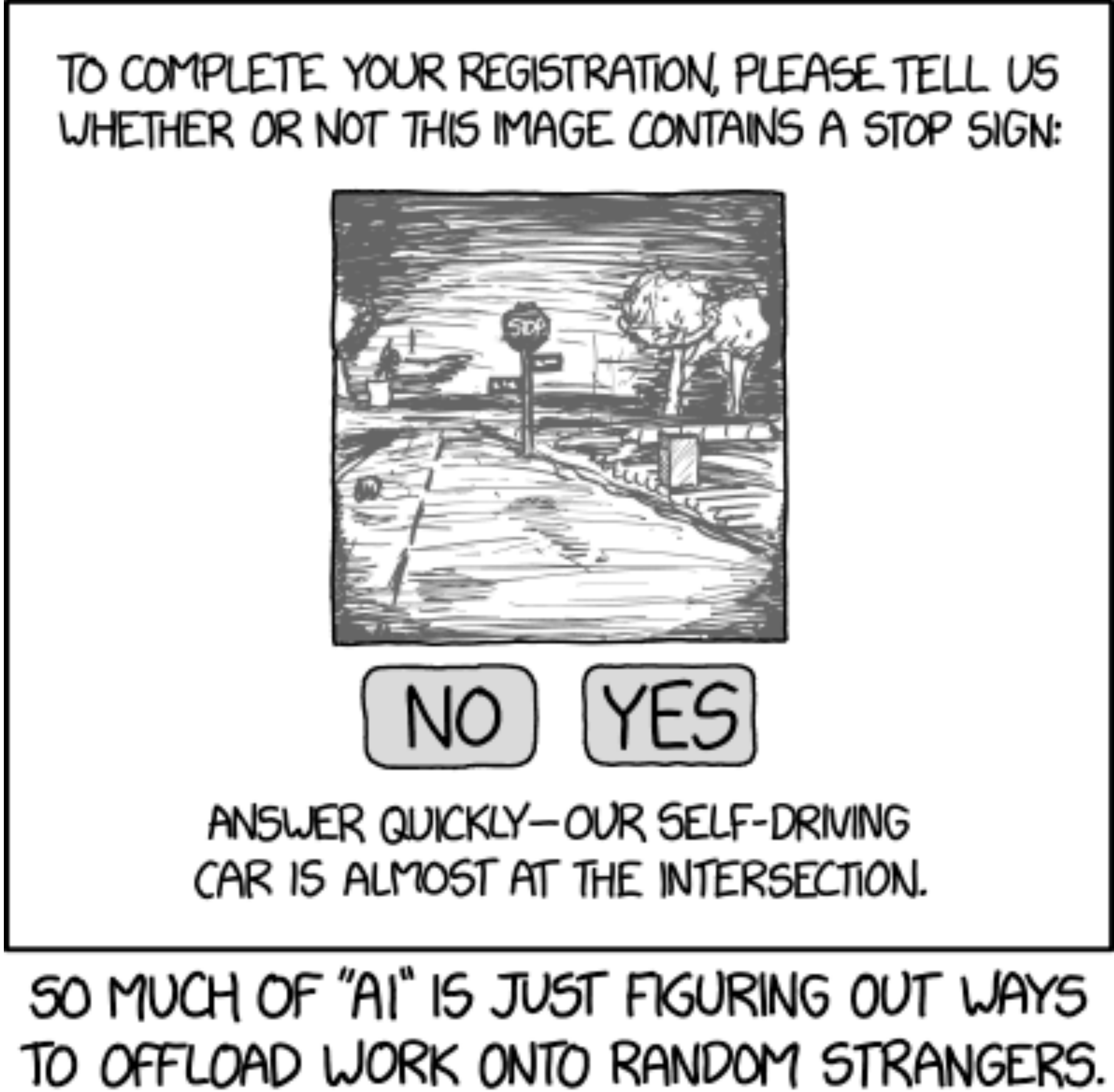
○ ОРГ
○ ПЕРС
○ ЛОК
● РАЗН

Коментар ✕

Додати фрагмент | Видалити | Перемістити | OK | Cancel

Delete this annotation.

Data-as-a-side-effect


TO COMPLETE YOUR REGISTRATION, PLEASE TELL US WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:
STOP
NO
YES
ANSWER QUICKLY—OUR SELF-DRIVING CAR IS ALMOST AT THE INTERSECTION.
SO MUCH OF "AI" IS JUST FIGURING OUT WAYS TO OFFLOAD WORK ONTO RANDOM STRANGERS.

Generating Language Data

- Potentially unlimited volume
- You control the parameters
- But! Artificial (is it representative?)

How? Take data you already have and replace words with synonyms, replace noun phrases with other noun phrases, come up with heuristic rules (your assumptions about how this data could look), etc.