University of Zurich[UZH]

Linguistic Research Infrastructure – Language Technology Group

A research infrastructure built on PostgreSQL

Johannes Graën & Jonathan Schaber

Friday 30th June, 2023

Outline

Introduction

The Linguistic Research Infrastructure ‘LiRI’

- infrastructure unit at the University of Zurich
- part of the Swiss Roadmap for Research Infrastructures
- in operation since 2020
- part of the Linguistic Center Zurich ‘LiZZ’
- offering services in the field of language data
- different groups (lab, language acquisition, statistics & ML, language technology)

The Language Technology Group

- Natural Language Processing (NLP)
- Application development
- IT infrastructure for the LiRI, LiZZ and customers (groups and projects at UZH)

Publicly available services

- Swissdox@LiRI – dataset compilation from 24m Swiss news articles
- LiRI Corpus Platform (LCP) – universal tool to query and inspect corpora

Us


Stefan Bircher
Tilia Ellendorff
Jean-Philippe Goldman
Johannes Graën
Daniel McDonald
Igor Mustač
Iuliia Parfenova
Nikolina Rajović
Klaus Rothenhäusler
Jonathan Schaber
Gerold Schneider
Bernard Schroffenegger
Ahmet Uluslu
Teodora Vuković
Jeremy Zehr

Services (Data Life Cycle)

Acquisition

- lab data
- web scraping
- corpus archives

Processing

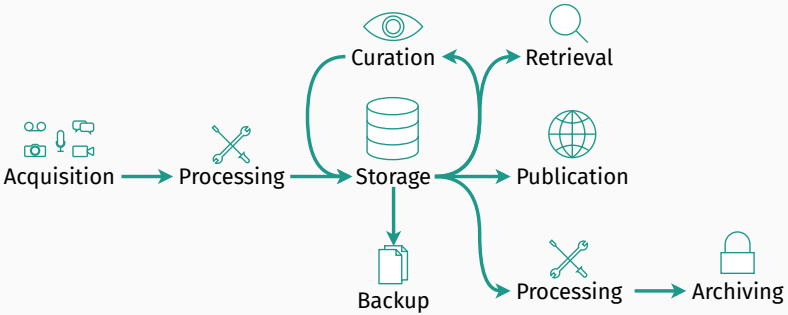
- conversion
- cleaning
- consolidation

Curation

- custom-made tools

Retrieval

- query engines



Publication

- headless web applications

Backup

- incremental Backups

Archiving

- long-term storage

Our use of PostgreSQL

We use PostgreSQL for all our tools

- that allow users to collaboratively work on data
- that allow users to query and/or explore complex or large data collections

We don’t use PostgreSQL for

- static data collection with defined access path
  (instead, we use JAMstack or tailer-made solutions)
- third-party applications with low requirements
  (we use MariaDB or SQLite instead)

Corpus Linguistics
(brief introduction)

A corpus (plural: corpora)

- collection of language samples
- often sampled according to some criterion (register, text sort, etc.)
- typically consist of text (but also audio, video, ...)
- reasonable sizes start around 1m tokens (words)
- used in corpus linguistics

Terminology

token smallest unit in a corpus

- words “hello world” → `[hello] [world]`
- punctuation “what’s” → `[what] ['s]`
- numbers “10 years” → `[10] [years]`

annotation inherent structures made explicit

- on token
- between tokens
- on other structures

→ we have tools for tokenization & annotation

Corpus annotations: parts of speech

Grouping of words into classes

- language specific
- Universal PoS tags *UPOS*

```
Thank you  ,     John Hume .
VERB  PRON PUNCT NOUN NOUN PUNCT
```

Corpus annotations: lemmatization

Mapping inflected word forms to their base forms

- some word classes are inflected
- better than stemming
  - *meeting* (Verb Part.) → stemmed: *meet*, lemma: *meet*
  - *meeting* (Noun) → stemmed: *meet*, lemma: *meeting*

and exemption from paying taxes when the occupational pension is redeemed .
and exemption from pay tax when the occupational pension be redeem .

Corpus annotations: morphological features

Inflected words can be described along PoS-specific dimensions

English verbs

	person	number	tense	mood	diathesis
(we) went	1	plural	simple past	indicative	active
(she) went	3	singular	simple past	indicative	active
grown	-	-	past participle	-	passive

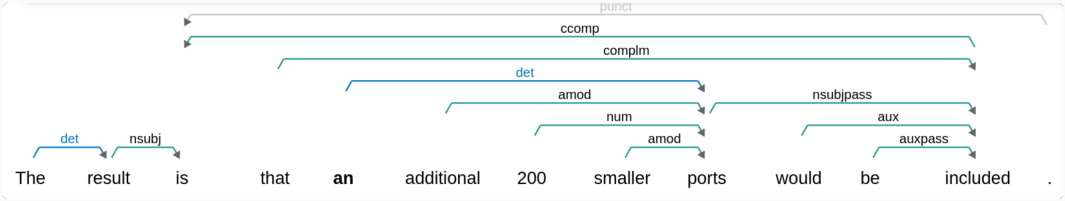
German nouns

	number	case
Hundes	singular	genitive
Freunden	plural	dative

Corpus annotations: dependencies

Syntactic framework capturing hierarchical relations between words

- directed links between 2 words (head $\rightarrow$ dependent)
- finite verb is root



Other frequent annotation layers

- coreference resolution
- named entity recognition (NER)
- sentiment analysis
- multimodal corpora
  - gestures
  - body positions
  - automated speech recognition ASR
- ...

Why Postgres?

- a relational database allows for complex, non-trivial queries (FOL)
- performance tuning (text-specific indices, etc.)
- scales well
- ACID for free

Challenges for linguistic query tools

1. performance
2. scaling (size & quantity)
3. usability (corpus query language: expressive vs. simple)

Our Corpus Platform

Goals

- facilitate corpus queries and exploration
- represent different types of corpora (modality, structure, annotation)
- deliver first results fast
- allow for complex queries – no postprocessing necessary

Our test corpora

- British National Corpus
  - written and spoken language from a wide range of sources between 1980s and 1993
  - newspapers, specialist periodicals and journals, academic books and popular fiction, etc.
  - more than 100 million tokens
- Open Subtitles (only the English part)
  - transcribed and translated subtitles for wide range of movies (since 1874)
  - usually short sentences (or fragments)
  - more than 3 billion tokens
- Europarl (de, en, es)
  - proceedings of the European Parliament between 1996 and 2011
  - high-quality translations, often longer and more complex sentence structures, parallel
  - more than 40 million tokens per language
- Tangram
  - video recording for interaction studies
  - various levels of annotation (manual and automatic)

British National Corpus

He stood up at once and began to stalk away but Mr Hellyer called to him.

Bawden, Nina. *Tortoise by candlelight*. London: Virago Press Ltd, 1989.

And I think they're ones that have a cartridge in.

31 conversations recorded by 'Martine' (PS0LK) between 12 and 20 March 1992 with 10 interlocutors.

They say that although many shops THINK they're providing for the disabled, the facilities are often inadequate.

[Central television news scripts]. Abingdon: Central TV, 1993.

It’s my wish you show your skill at a welcome feast for the sultan.

The 7th Voyage of Sinbad (1958)

Walker was a bad guy, and the more I find out about him, the happier I am he’s dead.

Body Heat (1981)

Of course, if it’s just an ordinary panther... it wouldn’t be big enough for a blanket.

Track of the Cat (1954)

German Die nächste Stufe wird die kommerzielle Nutzung sein, die nach unserer eigenen Richtlinie über die Patentierung biotechnischer Erfindungen zulässig ist.

English The next stage will be commercial exploitation, which our own bio-patenting directive allows for.

French La prochaine étape sera leur exploitation commerciale, ce qu'autorise notre propre directive relative à la protection juridique des inventions biotechnologiques.

September 6th 2000, Ahern, Nuala (Ireland), PoliticalGroup: Verts/ALE, OriginalLanguage: English

Annotations

Our text corpora have been annotated with

1. lemmas (cf. stemming)
2. part-of-speech tags (different tag sets)
3. syntactic dependencies

In addition, we have

- alignment (on documents, segments and tokens) in Europarl
- start and end times for utterances and gestures in Tangram

Performance

Question
How can we maximize query performance if we don't have defined access paths?

1. partition the data into random chunks of increasing sizes
2. represent attributes as FTS vectors to help restricting the search space early on
3. use Nested Sets for syntactic trees
4. ... and range types for hierarchical structures

(1)

Partition the data into random chunks of increasing sizes

Application requirements

- users are typically (any type of) linguists
- query definition is an iterative process (you know that)
- often a first glimpse is sufficient to assess the results

Application design

- deliver query results based on a (random) subset = subcorpus
- results can be examples, statistical and collocation analyses
- keep some examples cached in the backend
- continue updating analyses until a sufficiently large subset has been examined
- continue with the rest of the corpus on user request

Approach

Preparation

1. we assign a random uuid to each sentence
2. ...and employ it for rule-based bitwise partitioning
3. that way, we separate a half, a quarter, an eights, ... of the data until we reach a partition size of approximately one million sentences

Application

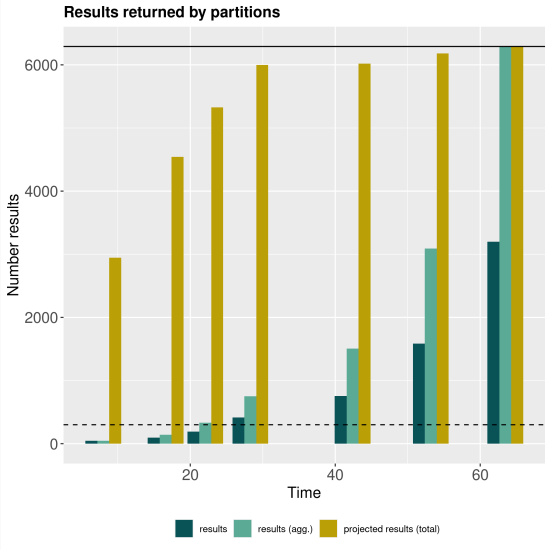
1. query the smallest partition
2. estimate hits for the remaining partitions based on results
3. continue with the partition that we expect to hold the missing number of results
4. repeat if necessary

Aggregated results

Query
We are looking for a determiner (a, the, ...) followed by an adjective followed by a noun that starts with “fr” and consists of at least five letters.

Corpus
Europarl (43m tokens)

Target
300 hits
(46 in first partition)



Examples

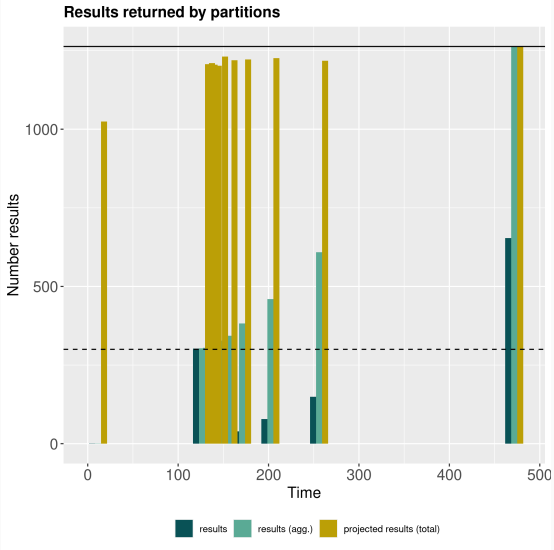
This fight will not curtail freedoms , it is a fight to defend	**the intrinsic freedom**	of every individual , namely the right to life and the right not to be …
… European Commission to play an active and firm role in defending	**the full freedom**	of movement .
… on the part of the public authorities , which we are , in laying down	**a legal framework**	with continuing to make use of the resources obtained as a result of …
… Commission considers that the proposal under consideration is not	**the appropriate framework**	for this discussion , which should take place within the context of …
we are not adopting	**a legislative framework**	to fight smoking , which continues to be the largest cause of death …
… upon which to base all arms export decisions , and to provide	**a well-defined framework**	for action , as a context for discussions of each case of arms transfer .
… for transport operations in the internal market , thereby creating	**a general framework**	for marketing products on the single European market .
… , some proposals) which came under the former third pillar (where	**the legal framework**	has changed substantially and they have therefore lapsed and must …
… to get to grips with the issues that I mentioned , we need to create	**a suitable framework**	for our future relations .
With regard to the sphere of application ,	**this legal framework**	must also apply to existing bilateral agreements .
… and nothing else , then that is when the Commission must establish	**a political framework**	for the European energy policy .
… much emphasis in the assistance we are giving on ensuring that	**the proper framework**	is established , in particular in relation to banking and the finance …
I voted in favour of this report aimed at providing	**a better framework**	for the activities of interest representatives in the EU .

Aggregated results (2)

Query
We are looking for an adverb followed by an adjective followed by a noun with the lemma “tree”.

Corpus
OpenSubtitles (3.1b tokens)

Target
300 hits
(first one in third partition)



Examples

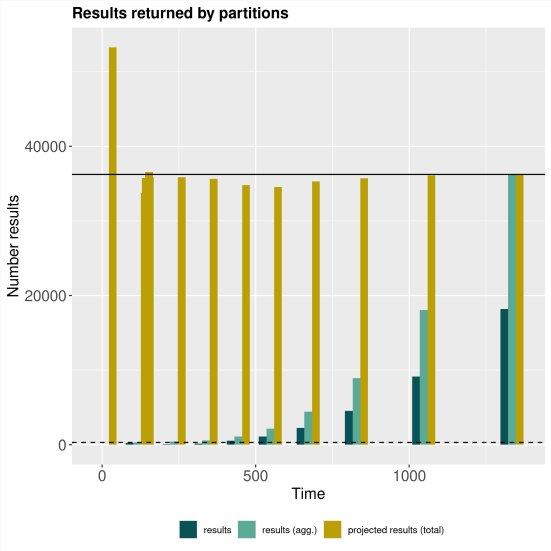
And he must have spent months on literally the	**most boring tree**	house in the world .
	How many trees	gave their lives for that ?
That is a	**perfectly salvageable tree**	.
And before our eyes a	**mathematically perfect tree**	appears .
	How many trees	are there ?
I shall ask Jackson to plant some	**very tall trees**	.
There was one	**very fine tree**	there .

Aggregated results (3)

Query
We are looking for the sequence “able to” followed by a verb that governs a preposition.

Corpus
Europarl (43m tokens)

Target
300 hits
(13 in first partition)



Examples

The EU must come up with a solution and be	**able to set**	an example	**for**	these countries .
… could be obtained in energy efficiency we should be	**able to achieve**	two thirds	**by**	2010 , that is an improvement of about 12 % in energy …
… railway passengers , who , from 2010 or 2012 will be	**able to travel**	more easily	**in**	their own countries and throughout the European Union .
… standards that are finally generally accepted , it should be	**able to succeed**	,	**without**	becoming a substitute for labour regulations or collective …
In this way , they have been	**able to create**	prosperity	**in**	a very much better way than they would have done by …
… which have already been adopted , on which we were only	**able to express**	an opinion	**after**	the event , despite the coordination method , which it …
This occupation , too , will thus be	**able to benefit**	fully	**from**	the terms of the internal market , and hauliers will enjoy …
Young people will be	**able to take**	part	**in**	sporting pursuits only under the supervision of their …
As I was	**able to point**	out	**in**	my speech during the debate , this amendment precisely …
… adopts its resolution today , the Commission will be	**able to adopt**	its revised communication	**in**	the middle of April , as has already been mentioned .
… of long-term resident is granted will the beneficiary be	**able to enjoy**	equal treatment	**in**	comparison with the nationals of Member States in …

(2)

Represent attributes as FTS vectors
to help restricting the search space early on

Vector representation

- prefixes for different attributes (word form, lemma, part-of-speech tag, ...)
- ‘overload’ positional information with all attributes present
- manual definition of vector (no stemming, no stop word removal!)

Example

	form	lemma	part of speech
1	Social	social	ADJ
2	human	human	ADJ
3	rights	right	NOUN
4	complement	complement	VERB
5	the	the	DET
6	traditional	traditional	ADJ
7	liberal	liberal	ADJ
8	freedoms	freedom	NOUN
9	.	.	.

Use vectors to restrict search space

Prefixes used: [1] = word form, [2] = lemma, [3] = part-of-speech

- literal sequence of “an Old Man”: `[1]an <1> [1]Old <1> [1]Man`
- allow for any noun (“and Old XX”): `[1]an <1> [1]Old <1> [3]NOUN`
- accept also plural forms (“the old men”): `[2]the <1> [2]old <1> [2]man`
- any noun phrase with a single determiner and adjective: `[3]DET <1> [3]ADJ <1> [3]NOUN`

Use vectors – positional overloading

- Multiple words can be specified on the same location!
- Information on that feature is scarce in the documentation.
  A special case that's sometimes useful is that <0> can be used to require that two patterns match the same word.

(The only mention of `<0>` for a zero distance.)

Hands-on example

```
db=> SELECT 'This is a test'::tsvector;
        tsvector
------------------------
 'This' 'a' 'is' 'test'
```

Compare to using the `to_tsvector` function:

```
db=> SELECT to_tsvector('This is a test');
 to_tsvector
-------------
 'test':4
```

Hands-on example – overloading positions

```
db=> SELECT 'This:1 is:2 a:3 an:3 test:4 example:4'::tsvector;
                      tsvector
-----------------------------------------------------
 'This':1 'a':3 'an':3 'example':4 'is':2 'test':4
```

Hands-on example – “a test”

```
db=> SELECT 'This:1 is:2 a:3 an:3 test:4 example:4'::tsvector
db-> @@ 'a <-> test';
 ?column?
----------
 t
```

Hands-on example – “an example”

```
db=> SELECT 'This:1 is:2 a:3 an:3 test:4 example:4'::tsvector
db-> @@ 'an <-> example';
 ?column?
----------
 t
```

Hands-on example – overloading positions

```
db=> SELECT
db-> 'The:1 Commission:2 then:3 found:4 a:5 few:6 more:7 reports:8 .:9
db-> DET:1 NOUN:2 ADV:3 VERB:4 DET:5 ADJ:6 ADJ:7 NOUN:8 .:9'::tsvector
db-> @@ 'ADJ <1> ADJ <1> NOUN';
 ?column?
----------
 t
```

Hands-on example – overloading positions

```
db=> SELECT
db-> 'The:1 Commission:2 then:3 found:4 a:5 few:6 more:7 reports:8 .:9
db-> DET:1 NOUN:2 ADV:3 VERB:4 DET:5 ADJ:6 ADJ:7 NOUN:8 .:9'::tsvector
db-> @@ 'a <1> few <1> ADJ';
 ?column?
----------
 t
```

Real-world example

Prefixes used: [1] = word form, [2] = lemma, [3] = part-of-speech

```
'¹.':31 '¹So':1 '¹a':10,28 '¹accident':30 '¹event':26 '¹give':14
'¹greatest':16 '¹have':7 '¹important':5 '¹in':8,24 '¹is':3 '¹it':2
'¹members':20 '¹nuclear':29 '¹of':21,27 '¹place':9 '¹possible':17
'¹public':23 '¹reassurance':18 '¹system':11 '¹the':15,22,25
'¹to':6,19 '¹very':4 '¹which':12 '¹will':13
'².':31 '²a':10,28 '²accident':30 '²be':3 '²event':26 '²give':14
'²great':16 '²have':7 '²important':5 '²in':8,24 '²it':2 '²member':20
'²nuclear':29 '²of':21,27 '²place':9 '²possible':17 '²public':23
'²reassurance':18 '²so':1 '²system':11 '²the':15,22,25 '²to':6,19
'²very':4 '²which':12 '²will':13
'³.':31 '³ADJ':5,16,17,29 '³ADP':8,21,24,27 '³ADV':1,4
'³DET':10,12,15,22,25,28 '³NOUN':9,11,18,20,23,26,30 '³PRON':2
'³PRT':6,19 '³VERB':3,7,13,14
```

Performance gain

- a vector search prior to the actual search can restrict the search space a lot
- 1630 occurrences of “Big Mac” in OpenSubtitles retrieved in < 50ms (>400m vectors)
- but: 51k occurrences of “book” as a verb takes 50s on the same partition
- some vector searches are futile, e.g. looking for sentences that comprise a verb

Partitioned vectors

Size of a partition corresponds to:

$$\frac{N}{2^p}$$

- N = number of tokens in corpus
- p = enumerated partition

Vector query was:

```
'3ADV <1> 3ADJ <1> 2tree'
```

p	size (in m)	query times (ms)
12	0.2	32.487
11	0.4	65.457
10	0.8	134.037
9	1.7	265.249
8	3.3	534.265
7	6.7	1076.642
6	13	2147.029
5	27	4282.480
4	53	8630.625
3	107	17339.286
2	213	34438.617
1	427	70147.207
0	854	144676.546

More advanced queries

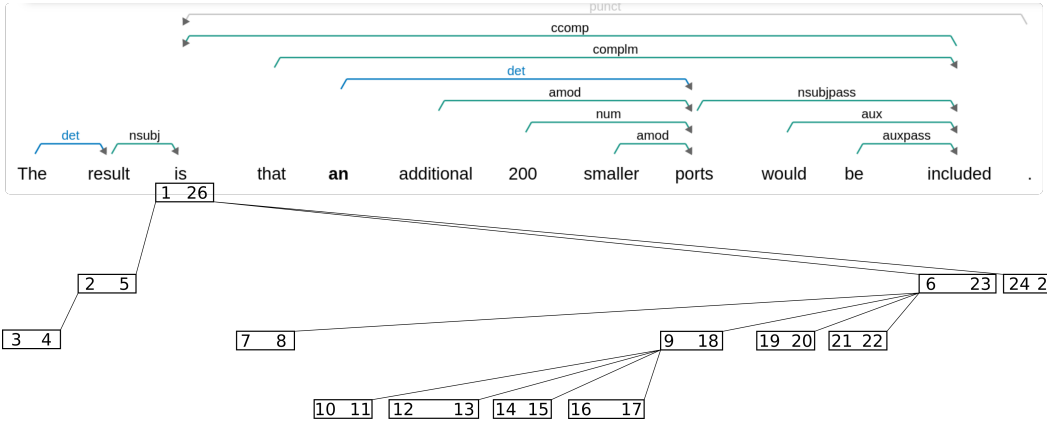
- repetition patterns in queries are translated to partial vector queries:
  “determiner followed by **two or more** adjectives followed by a noun” ⇒
  `[3]DET <1> [3]ADJ <1> [3]ADJ & [3]ADJ <1> [3]ADJ <1> [3]NOUN`
- optional repetitions (zero or more) can only be used if an upper limit is set:
  “determiner **optionally** followed by an adjective followed by a noun” ⇒
  `[3]DET <1> ([3]ADJ <1> [3]NOUN | [3]NOUN)`
- any logical operators can be taken over to the vector query:
  “determiner followed by **either** an adjective **or** an adverb followed by a noun” ⇒
  `[3]DET <1> ([3]ADJ | [3]ADV) <1> [3]NOUN`

(3)
Use Nested Sets for syntactic trees

Dependencies as Nested Set

- syntactic dependencies form tree structure
- different ways to relationally represent
  - adjacancy list
  - closure table
  - nested set
- often simplify otherwise recursive queries to *BETWEEN* constraints

Dependencies as Nested Set



Dependencies as Nested Set

token	left	right	dep-rel
The	3	4	det
result	2	5	nsubj
is	1	26	
that	7	8	complm
an	10	11	det
additional	12	13	amod
200	14	15	num
smaller	16	17	amod
ports	9	18	nsubjpass
would	19	20	aux
be	21	22	auxpass
included	6	23	ccomp
.	24	25	punct

token	left	right	dep-rel
The	3	4	det
result	2	5	nsubj
is	1	26	
that	7	8	complm
an	10	11	det
additional	12	13	amod
200	14	15	num
smaller	16	17	amod
ports	9	18	nsubjpass
would	19	20	aux
be	21	22	auxpass
included	6	23	ccomp
.	24	25	punct

(4)
Use range types for hierarchical structures

Range Types

- range types *int4range*, *int8range*
- operators

  A <@ B A is contained by B
  A @> B A contains B
  A && B A and B overlap
  A << b A is strictly left of B

  ...

- indexable with GiST
- whole corpus is one virtual character stream, entities have a start & end character
- well-suited to skip hierarchical layers
  token *<@* segment *<@* paragraph *<@* chapter *<@* book

Plans

Future plans

Things we would like to try:

- extract positional information from vector queries
- extend vector queries to allow for repetition operators (like regexp on lexemes)
- use grouping sets to calculate statistics on the fly
- use multiple cursors (for raw data and analyses)

Questions?