Navigating tenths of gigabytes of **English** text **entangled** with its **definition** and **structure** efficiently **on a laptop**.

# The Tangle: A Connected Dataset of English Grammar and Curated Text

Jacques Basaldúa
Lead author of the Jazz platform
Senior Data Scientist @ BBVA Data & Analytics

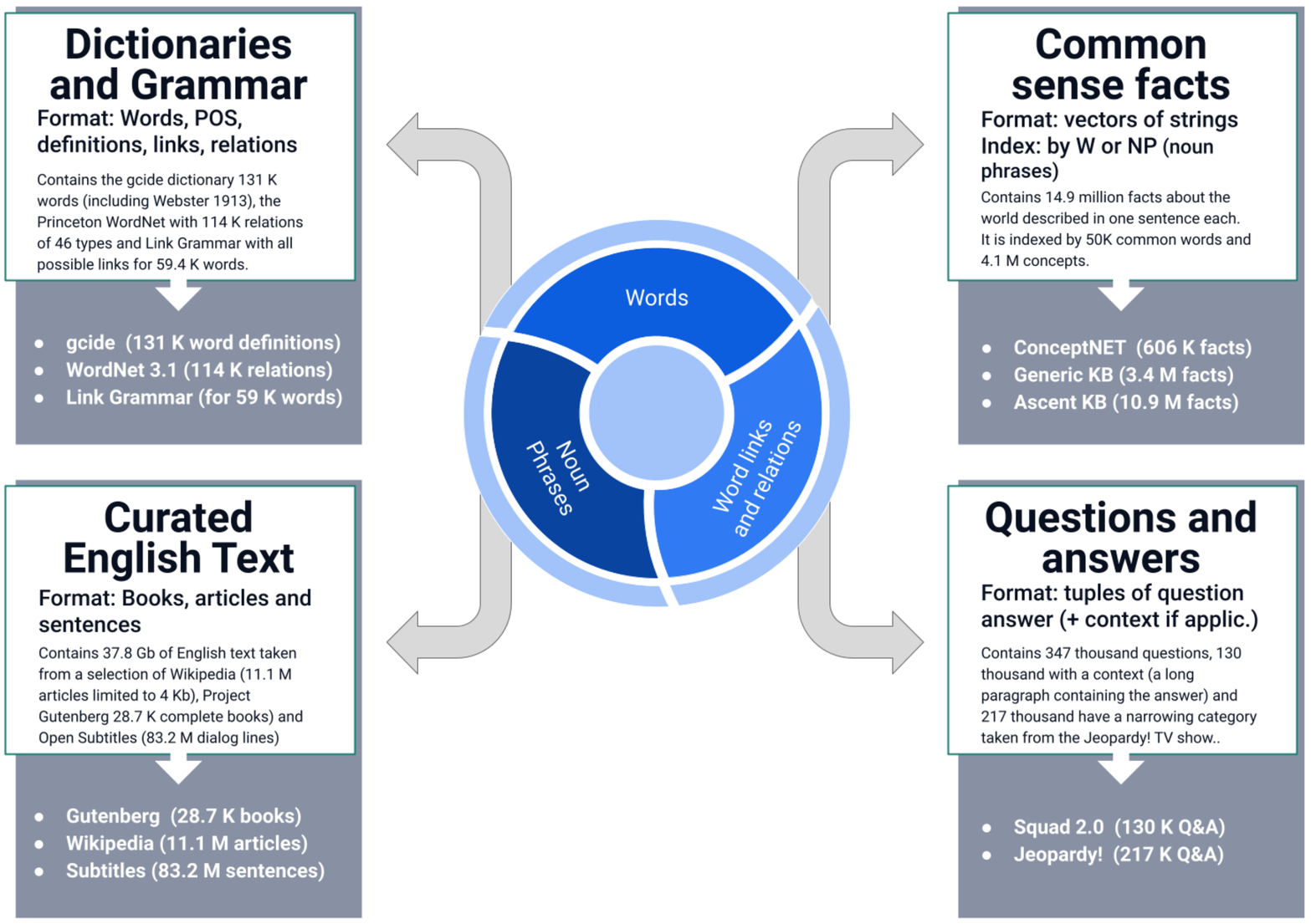
## 1 Introduction

- The recent success of NLP methods using vector embeddings and transformers has shadowed a lot of the previous research that was grounded on solid linguistics, dictionaries and grammatical resources. While modern deep learning platforms are very efficient at computing vectors and transformers, they are not well suited for including resources written in text such as dictionaries, grammars and fact databases as part of the language model. This **Language in Language** (LiL) approach requires both curated datasets and technology that enables efficient navigation through the graph of "entangled English".
- This is where **The Tangle** jumps in. The Tangle is a curated collection of public datasets persisted inside a highly efficient data server, **Jazz**. Using the native platform language, C++, the connected language graph of words and phrases can be navigated at speeds of millions of edges per thread per second on a laptop using pre-computed indices. Since the server is an http server, it is easy to connect from any language, at least for exploration purposes.
- The Tangle is fundamental for our own research. It is released for the first time with Jazz 0.5.2 (November 2021) and will be improved and released with major Jazz releases. It focuses on quality rather than quantity, but is still large enough to hold the English Wikipedia and the Gutenberg Project on a laptop with around 50 Gb of disk space and small RAM requirements.
- The use of The Tangle is completely open ended, it is not a challenge, just text in a platform that enables walking the connections between words, concepts or noun phrases and text resources very efficiently. It makes reinforcement learning, active learning and lifelong learning ideas that need fast iteration possible. It will be an essential component in researching language as code using **formal fields** by our team.

## 2 Highlights

- Word constituents: *prefixes, suffixes, syllables*.
- Different word lists with distributions
- Definitions of the words
- Word links (as in Link Grammar)
- Definitions of the word links
- Word relations (as in Princeton WordNet)
- Definitions of the word relations
- Noun Phrases
- Wikipedia page titles
- English Wikipedia (11 M articles of 4Kb max)
- Selection of Project Gutenberg including masterpieces
- Questions and Answers datasets with and without context
- List of facts for commonsense reasoning
- Jazz server runs with a single **docker run**
- Serious C++ projects use **jazz elements** API
- Example Python bindings available via **pip install**

## 3 Diagram




## 4 More on Jazz

- Jazz 0.1.7 was released as **Open Source** by BBVA, end of 2017.
- Almost 3000 commits later, Jazz 0.5.2 is **the latest version**.
- It is the first release that can be downloaded with **The Tangle** included.
- Jazz implements **data blocks**: tensors, kinds and tuples.
- Jazz provides support for **communicating** and **storing** blocks.
- Jazz's persisted storage is based on **LMDB**.
- Jazz has **in RAM** block storage using: queues, trees, deques, ...
- Jazz can communicate blocks over: the file system, http, zeroMQ and bash.
- Everything is available through a **C++ API**.
- Besides C++, Jazz is an http server exposing everything via an **http API**.
- This allows access from any language, Python examples included.
- The http server can also deliver **apps stored in Jazz**.
- Apps are written with any **front framework**: Angular, Ionic, Dart, Flutter, ...
- Jazz 0.6.x (under development) provides backend code execution.
- Jazz 0.6.x will provide **language tools** (many already existing).
- Jazz 0.7.x (PoC available) will implement code generation algorithms.
- The **main focus** of the platform is **language technology**.

## 5 Large text corpora example

These code examples assume:

- You have succesfully installed docker and are running a Jazz 0.5.2 TNG (The Tangle) server
- You have succesfull installed `thetangle` python package

```
from thetangle import TangleExplorer as te
```

```
# TangleExplorer is just syntatic sugar for how to use an index to get get a slice to filter some persisted block.

# If you plan to use The Tangle from the http server, you have to understand:
#  - How the API works:         https://kaalam.github.io/jazz_reference/api_ref_intro.html
#  - How the tangle is stored:  https://kaalam.github.io/jazz_reference/reference_docker_tangle_server.html
```

```
# The Wikipedia Index is persisted, we neew to load it to a Jazz volatile index

te.load_new('//index/my_idx', '//lmdb/en_wiki_2021-10-01/titles')
```

```
True
```

```
te.get('Hercule Poirot', index = '//index/my_idx')
```

```
"ref is a fictional belgian detective created by british writer agatha christie. poirot is one of christie's most famous and long-running characters
novels, 2 plays '' black coffee play black coffee '' and '' alibi play alibi '', and more than 50 short stories published between 1920 and 1975. poi
rayed on radio, in film and on television by various actors, including austin trevor, john moffatt actor john moffatt, albert finney, peter ustinov,
andall, alfred molina, orson welles, david suchet, kenneth branagh, and john malkovich. overview influences poirot's name was derived from two othe
ives of the time marie belloc lowndes ' hercule popeau and frank howel evans ' monsieur poiret, a retired belgian police officer living in london.
ttp www.litencyc.com php speople.php?rec true uid 5054 title agatha christie 1890 1976 access-date 6 september 2006 first chris last willis locatio
tan university ref a more obvious influence on the early poirot stories is that of arthur conan doyle. in ''an autobiography'', christie states, i
in the sherlock holmes tradition eccentric detective, stooge assistant, with a inspector lestrade lestrade -type scotland yard detective, inspector
ced as the introduction to harvnb christie 2013 ref for his part, conan doyle acknowledged basing his detective stories on the model of edgar allan
dupin and his anonymous narrator, and basing his character sherlock holmes on joseph bell, who in his use of wikt ratiocination ratiocination prefi
iance on his little grey cells. poirot also bears a striking resemblance to a. e. w. mason 's fictional detective inspector hanaud of the french s
eared in the 1910 novel '' at the villa rose novel at the villa rose '' and predates the first poirot novel by 10 years. christie's poirot was clea
her early development of the detective in her first book, written in 1916 and published in 1920. belgium's occupation by germany during world war i
ble explanation of why such a skilled detective would be available to solve mysteries at an english country house. sfn christie 1939 at the time of
g, it was considered patriotic to express sympathy towards the belgians, ref cite book author horace cornelius peterson title propaganda for war. t
t american neutrality, 1914 1917 url https books.google.com books?id pkjbnqeacaaj year 1968 publisher kennikat isbn 9780804603652 ref since the inva
ntry had constituted britain's '' casus belli '' for entering world war i, and british wartime propaganda emphasised the rape of belgium."
```

## 6 Common sense example

```
from thetangle import TangleExplorer as te
```

```
# The Ascent Index is persisted, we neew to load it to a Jazz volatile index

te.load_new('//index/my_idx', '//lmdb/ascent/idx_np')
```

```
True
```

```
te.get('absolute value', index = '//index/my_idx')
```

```
['a maximum distortion of absolute value is usually the acceptable limit',
 'a positive absolute value could represent either a positive or a negative original value',
 'a probability amplitude is that thing that you take the square of well absolute value squared to get probability',
 'a quick word on notation the absolute value of a number is indicated in writing by putting the number between a pair of vertical bars',
 'a realvalued function v on a field f is called an absolute value also a modulus magnitude value or valuation if it satisfies the following four axion
 'a slope with a greater absolute value indicates a steeper line',
 'a stronger absolute value indicates a stronger linear relationship between the two variables',
 'abs absolute value distance from zero of a value or expression sign sign or of a value or expression fact factorial function',
 'absolute value absolute value can be a bit confusing for students who are just learning about signed numbers',
 'absolute value absolute value describes the distance of a number on the number line from without considering which direction from zero the number lie
 'absolute value also known as an intrinsic value refers to a business valuation method that uses discounted cash flow dcf analysis to determine a comp
worth',
 'absolute value can be explored both numerically and graphically',
 'absolute value can be expressed as a function fx x x if x is positive x if x is negative',
 'absolute value changes the value of the expression within them',
 'absolute value describes the magnitude of a number or the distance between points but it strips out information on the sign of the number or the dire
istance',
 'absolute value gives distance but discards information about direction',
 'absolute value has a specific notation in mathematical expressions',
 'absolute value has several properties some of which we have encountered earlier ill omit the repetitive for all expression that should precede each c
es',
 'absolute value in excel can be calculated using abs function which is available under the category of math and trig in insert function',
 'absolute value in excel is very simple and easy to use',
 'absolute value is a distinct mathematical principle that states a value without regard to its status as a positive or negative value',
 'absolute value is a helpful concept when we are only interested in the size of the difference between two numbers',
 'absolute value is a mathematical function that takes the positive version of whatever number is inside the absolute value signs which are drawn as tw
s',
 'absolute value is a numbers distance from zero or its nonnegative value',
 'absolute value is a term used in mathematics to indicate the distance of a point or number from the origin zero point of a number line or coordinate
 'absolute value is also known as magnitude',
 'absolute value is always positive or zero and a positive absolute value could result from either a positive or a negative original value',
 'absolute value is an easy concept absolute value of a nonnegative number is the number itself absolute value of a negative number is the opposite of
 'absolute value is calculated using the formula given below',
 'absolute value is defined as',
```

## 7 Questions and answers example

```
from thetangle import TangleExplorer as te
```

```
te.get('//lmdb/jeopardy_category', rows = [127])
```

```
'i lads'
```

```
te.get('//lmdb/jeopardy_question', rows = [127])
```

```
'czar at 17, he was famous for extraordinary sadism cruelty, even as a boy'
```

```
te.get('//lmdb/jeopardy_answer', rows = [127])
```

```
'ivan the terrible'
```

```
te.get('//lmdb/squad20_question', rows = [18912])
```

```
'in addition to much of the state of arizona, what u.s. state does not ever change their clocks for dst?'
```

```
te.get('//lmdb/squad20_question_ctx', rows = [18912])
```

```
'3196'
```

```
te.get('//lmdb/squad20_context', rows = [3196])
```

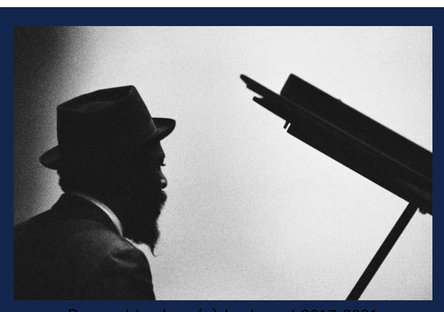
```
'coordination strategies differ when adjacent time zones shift clocks. the european union shifts all at once, at 01 00 utc or 02 00 cet or 03 00 eet f
tern european time is always one hour ahead of central european time. most of north america shifts at 02 00 local time, so its zones do not shift at t
r example, mountain time is temporarily for one hour zero hours ahead of pacific time, instead of one hour ahead, in the autumn and two hours, instead
of pacific time in the spring. in the past, australian districts went even further and did not always agree on start and end dates for example, in 200
rving areas shifted clocks forward on october 5 but western australia shifted on october 26. in some cases only part of a country shifts for example,
ii and most of arizona do not observe dst.'
```

```
te.get('//lmdb/squad20_answer', rows = [18912])
```

```
'hawaii'
```

Scan me

← Find evertything about The Tangle here!



Powered by Jazz (c) kaalam.ai 2017-2021