

Introducing *Comparapedia*

A new resource for Corpus-Based Translation Studies

S. Bernardini¹, S. Castagnoli¹, A. Ferraresi^{1,2}
F. Gaspari¹, E. Zanchetta¹

1. University of Bologna, Italy
2. University of Naples "Federico II", Italy



Background

Web corpora for translators @ Forlì

- **The WaCky! way**

- very large (>1 billion words) corpora
- built by crawling + “cleaning” data *post-hoc*
- *general-purpose* corpora for multiple languages: DE, EN, IT, FR

- **The BootCaT approach**

- tool to automatically build *specialised* corpora
- requires small sets of domain-specific *seeds*
- *ad hoc* corpora, usually small(-ish), for all European languages

A (free) user-friendly interface is now available (*beta version*)

Background

Why a Wikipedia corpus?

- Opportunity
 - lots of text, multilingual coverage, convenient format (xml, Wikipedia dumps), no copyright issues
- Practical/didactic interest
 - translators use Wikipedia as a source of factual information but
 - web format does not allow sophisticated linguistic queries
- Theoretical/descriptive interest
 - linked Wikipedia articles
 - independent entries in languages A and B
 - ST in lang. A and TT in lang. B (or two translations from lang. C)
 - ST and a heavily edited TT
 - how does our traditional notion of translation relate to collaborative web-based multilingual text production?

Turning Wikipedia into a *comparable* corpus

Corpus structure (IT/EN but replicable)

1. Two large, independent **monolingual corpora**
 - all of Wikipedia IT + all of Wikipedia EN
2. A smaller **comparable corpus**
 - all entries available both in IT and EN
3. A (much) smaller set of **parallel segments**
 - Translation Memory style
 - 1:1 matches only
 - linked to whole texts in the comparable corpus providing browsable co-texts

What we aim for

- A corpus
 - consisting of all explicitly linked bi-articles (in Italian and English)
 - allowing browsing of article pairs and
 - on-the-fly building of thematic subcorpora
- Guidelines and tools for others to replicate the procedure
 - for other language pairs
 - for future dumps (“monitor” Wikipedia corpus?)

Our starting point

```
<page>
<title>Ormskirk</title>
<id>296301</id>
<revision>
  <id>349597166</id>
  <timestamp>2010-03-13T11:37:29Z</timestamp>
  <contributor>
    <username>Dr Greg</username>
    <id>847224</id>
  </contributor>
  <comment>not much point in specifying distance to 100 m precision!</comment>
  <text xml:space="preserve">{{Infobox UK place
| country           = England
| latitude          = 53.5700
| longitude         = -2.8827
| official_name     = Ormskirk
| population        = 23,392
| shire_district    = [[West Lancashire]]
| shire_county      = [[Lancashire]]
| region            = North West England
| constituency_westminster = [[West Lancashire (UK Parliament constituency)|West Lancashire]]
| post_town         = ORMSKIRK
| postcode_district = L39
| postcode_area     = L
| dial_code         = 01695
| os_grid_reference = SD415085
| static_image      = [[Image:Ormskirk market.JPG|240px]]
| static_image_caption = &lt;small&gt;Market day in Ormskirk&lt;/small&gt;
}}
&lt;!--Start of article--&gt;
'''Ormskirk''' is a [[market town]] in [[West Lancashire|West]] [[Lancashire]], [[England]]. It is situated {{convert|13|mi|km|0}} north of [[Liverpool]] city centre, {{convert|11|mi|km|0}} northwest of [[St Helens, Merseyside|St Helens]], {{convert|9|mi|km|0}} southeast of [[Southport]] and {{convert|15|mi|km|0}} southwest of [[Preston]].

==Geography and administration==
Ormskirk lies on sloping ground on the side of a ridge, whose highest point is {{convert|68|m|ft|0}} above sea-level, at the centre of the [[West Lancashire Coastal Plain|West Lancashire Plain]],&lt;ref name=gillibrand Townships: Ormskirk>[http://www.british-history.ac.uk/report.asp?compid=41331&amp;strquery=gillibrand Townships: Ormskirk], British History Online&lt;/ref> and has been described as a &quot;planned borough&quot;; laid out in the thirteenth century.&lt;ref name=assessment>[http://www.lancashire.gov.uk/environment/archaeologyandheritage/historictowns/OrmskirkComplete_LowRes.pdf], Ormskirk
```

In practice...

1. Download Wikipedia dumps (18/03/10)
2. Extract XML files
3. Keep
 - references to entries in other languages
 - categories
4. Clean texts of markup and boilerplate (using *WikiExtractor*)

In practice (cont'd)

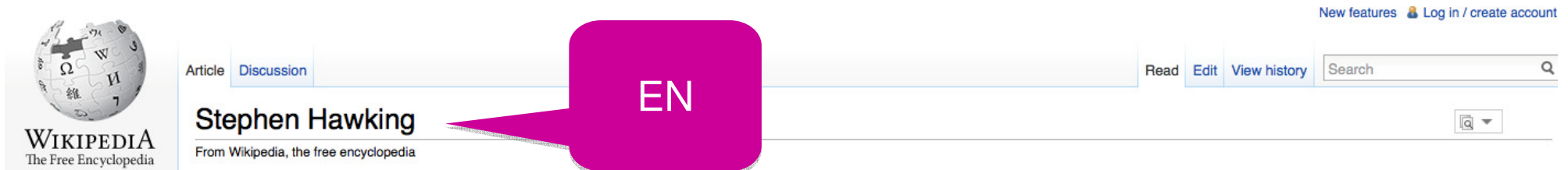
5. Only keep articles with EN<=>IT link
6. Metadata:
 - text id (= article's title in lang. A)
 - text target (= matching article's title in lang. B)
 - categories
7. POS-tag and lemmatise (TreeTagger)
8. Index with the Corpus WorkBench
 - *Comparapedia* EN
 - *Comparapedia* IT

Aside: Categories

from Wikipedia to *Comparapedia*

- Original Wikipedia categories
 - inserted by humans
 - richer in EN than in IT
 - some work done in NLP to give them structure
 - YAGO; DBPEDIA; WIKINET
- Our “quick and dirty” approach
 - lowercase
 - keep only lexical words => keywords
 - sort in alphabetical order
 - migrate EN keywords to matching IT article

From categories to keywords



<text_keywords 1942 20th-century 21st-century academics academy adams albans albert alumni applied arts astronomers astronomical births british caius calculating cambridge college commanders companions copley cosmologists department disease einstein empire english fellows former freedom gold gonville hall hertfordshire honorary honour laureates living lucasian mathematics medal members motor national neuron order oxford people philosophers physicists physics pontifical presidential prize prodigies professors pupils recipients religious royal school science sciences skeptics society st theoretical trinity university wolf writers>



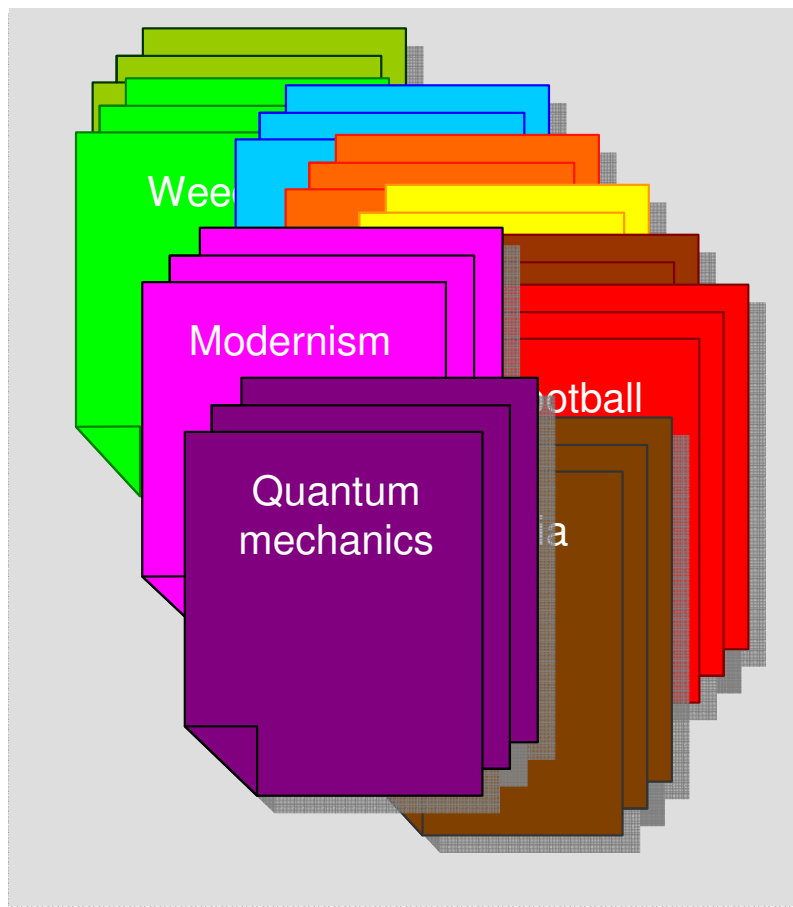
Categorie: [Matematici britannici](#) | [Astrofisici britannici](#) | [Nati nel 1942](#) | [Nati l'8 gennaio](#) | [Bambini prodigio](#) | [Fisici teorici](#) | [Saggisti britannici](#) | [Divulgatori scientifici britannici](#) | [Membri della Royal Society](#) | [\[altre\]](#)

Quick corpus facts

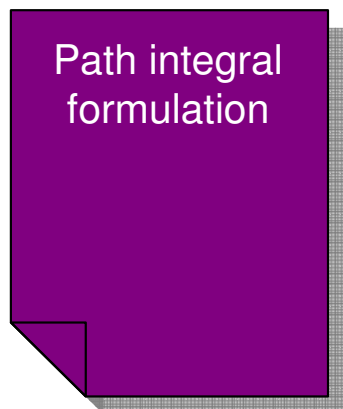
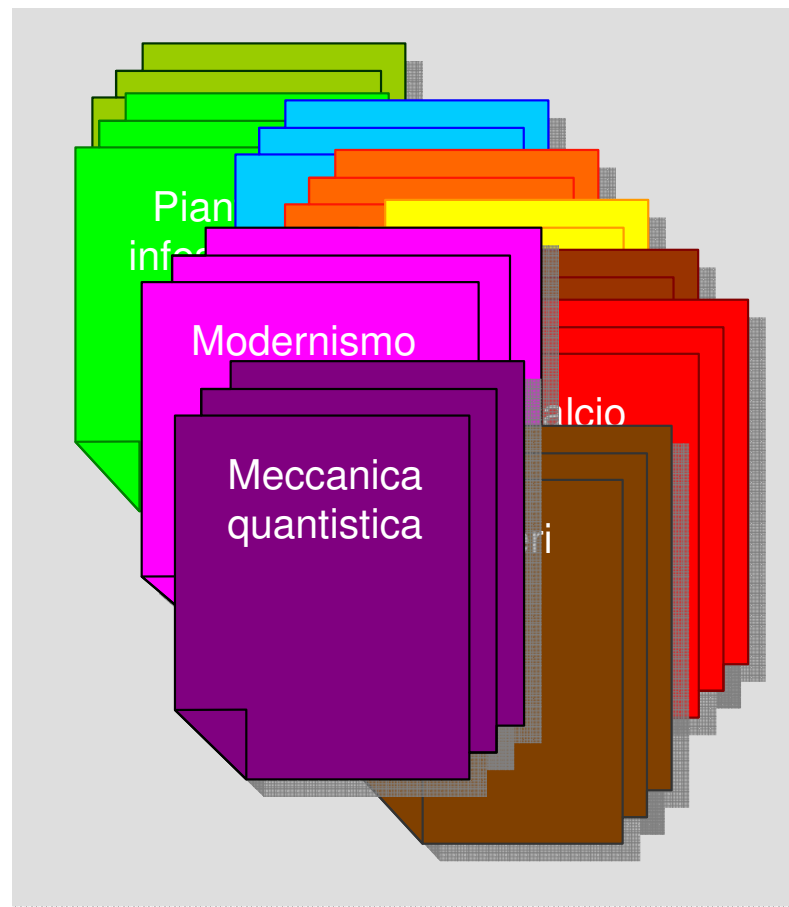
	Comparapedia EN	Comparapedia IT
Articles	426,273	426,057
Tokens	274,344,165	139,975,783

- Corpus structure – pseudo xml
 - `<text id="title" target="target_title" keywords="kw1 kw2 kwn">`
 - the actual text in vertical format
(positional attributes: word, pos, lemma)

Comparapedia EN



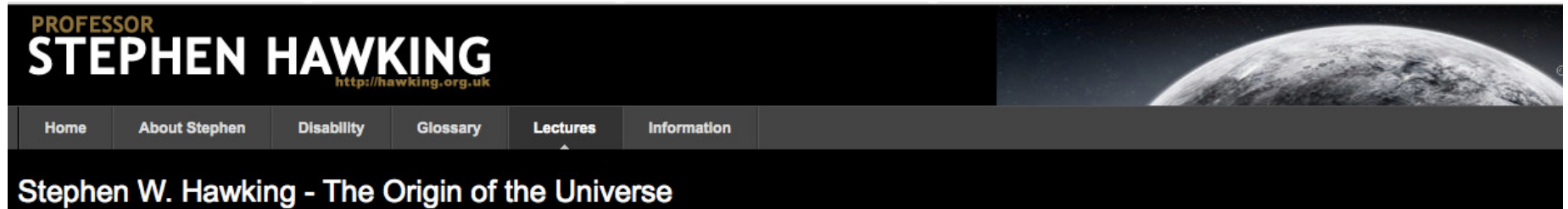
Comparapedia IT



Comparable subcorpora

Matching text pairs

An example...



“In order to understand the Origin of the universe, we need to combine the General Theory of **Relativity** with **quantum** theory. The best way of doing so seems to be to use **Feynman's** idea of a **sum over histories**. Richard Feynman was a colorful character, who played the bongo drums in a strip joint in Pasadena, and was a brilliant physicist at the California Institute of Technology. He proposed that a system got from a state A, to a state B, by every possible path or history. Each path or history has a certain amplitude or intensity, and the probability of the system going from A- to B, is given by adding up the amplitudes for each path. There will be a history in which the moon is made of blue cheese, but the amplitude is low, which is bad news for mice.”

“Sum over histories”

<text_id World>: for the <sum of human civilization living, specifically human experience, **history**>, or
<text_id Southwest Airlines>: a <sum which would have been the largest fine in the agencies **history**> - was
<text_id Britain's Got Talent>: figure <sum in what has been the biggest surprise in reality TV **history**>.
<text_id Yoruba people>: Itan is the term for the <sum total of all Yoruba myths, songs, **histories**>, and
<text_id Species>: can be <summed up insofar as that the BSC defines a species as a consequence of
manifest evolutionary "**history**" , while the PSC
<text_id Land of Punt>: majority of Egyptologists is <summed up by Ian Shaw from the Oxford **History**> of

Lemmas SUM &
HISTORY within <s>
in Comparapedia
EN
(total = 48 hits)

Same search in texts with
keywords
feynman | quantum | relativity
(total = 5 hits)

<text_id Feynman diagram>: amplitude as a weighted <sum of all possible **histories**> of the system
<text_id Feynman diagram>: for scattering is the <sum of each possible interaction **history**> over
<text_id Quantum mechanics>: mechanical amplitude is considered as a <sum over **histories**> between
<text_id Path integral formulation>: quantum mechanics, the "<sum over **histories**>" interpretation
<text_id Path integral formulation>: event is. The <sum over **histories**> method gives identical

The expression is a domain-specific term in English.
How about Italian?

“Somma sulle storie”?

- 26 hits from Google
- 1 hit from *Comparapedia* IT (domain: maths)
- BUT the **idea** of “sum over histories” is bound to be expressed (somehow) in the 3 Italian articles corresponding to English
 - Feynman diagram (*~ Diagramma di Feynman*)
 - Quantum mechanics (*~ Meccanica quantistica*)
 - Path integral formulation (*~ Integrale sui cammini*)

The matching Italian texts
become our micro-corpus

“Integrale sui cammini”





<text id=“Integrale sui cammini” target=“Path integral formulation”>: L'integrale sui cammini (o “path integral”) rappresenta una formulazione della meccanica quantistica che descrive la teoria quantistica generalizzando il principio di azione della meccanica classica . Esso rimpiazza **la classica nozione di una singola e unica storia di un dato sistema con una somma, o integrale funzionale, estesa a una infinità di possibili storie**, legate a infiniti modi di raggiungere una stessa configurazione quantistica, per il calcolo dell'ampiezza di probabilità. L'integrale sui cammini è stato sviluppato da Richard Feynman nel 1948.

- No lexicalised equivalent of “sum over histories” in Italian
- Either the term is paraphrased, or
- The more formal “*integrale sui cammini*” (=path integral) is used

The next steps

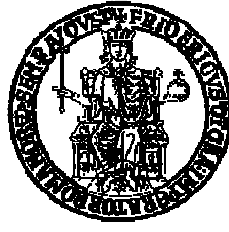
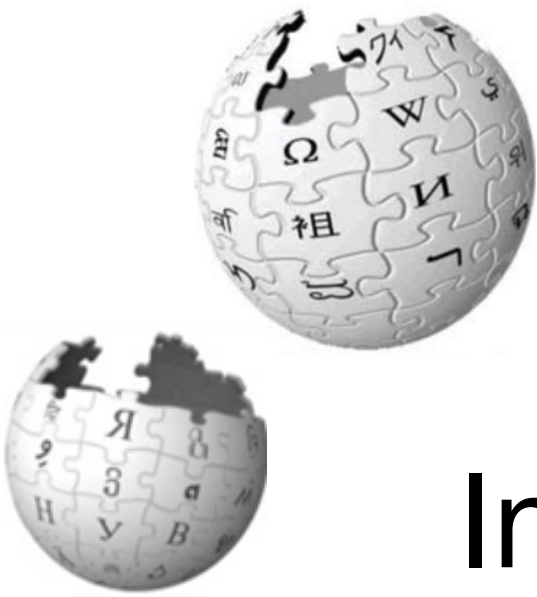
- Short term
 - leveraging work on Wikipedia-derived ontologies to make dynamic generation of specialised corpora more effective
- Longer term
 - work on the parallel dimension

Going parallel: prospects

	text id="Path integral formulation" target="Integrale sui cammini"	text id="Integrale sui cammini" target="Path integral formulation"
1. 	The "path integral formulation" of quantum mechanics is a description of quantum theory which generalizes the action principle of classical mechanics.	L'integrale sui cammini (o "path integral") rappresenta una formulazione della meccanica quantistica che descrive la teoria quantistica generalizzando il principio di azione della meccanica classica.
2. 	It replaces the classical notion of a single, unique trajectory for a system with a sum, or functional integral, over an infinity of possible trajectories to compute a quantum amplitude.	Esso rimpiazza la classica nozione di una singola e unica storia di un dato sistema con una somma, o integrale funzionale, estesa a una infinità di possibili storie, legate a infiniti modi di raggiungere una stessa configurazione quantistica, per il calcolo dell'ampiezza di probabilità.
3. 	The basic idea of the path integral formulation can be traced back to P. A. M. Dirac in his 1933 paper.	
4. 	The complete method was developed in 1948 by Richard Feynman.	L'integrale sui cammini è stato sviluppato da Richard Feynman nel 1948.

References

- BootCaT front-end: <http://bootcat.sslmit.unibo.it/>
- Corpus WorkBench: <http://cwb.sourceforge.net/>
- DBpedia: <http://dbpedia.org/>
- TreeTagger: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>
- WikiExtractor: http://medialab.di.unipi.it/wiki/Wikipedia_extractor
- WIKINET: Nastase, Strube, Börschinger, Zirn and Elghafari (2010) “WikiNet: A very large scale multi-lingual concept network”. *Proceedings of LREC 2010*
- YAGO: Suchanek, Kasneci and Weikum (2007) “Yago - A Core of Semantic Knowledge”. *Proceedings of 16th World Wide Web conference (WWW 2007)*



Introducing *Comparapedia*

THANK YOU!

silvia | scastagnoli | adriano | fgaspari | eros

@sslmit.unibo.it





silvia | scastagnoli | adriano | fgaspari | eros
@sslmit.unibo.it

