Interwiki: to correct or not to correct?

Łukasz Bolikowski

ICM, University of Warsaw + IBS, Polish Academy of Science

Wikimedia Polska 2009, 1-3 May 2009

Interlanguage links in Wikipedia

Interlanguage link, or **interwiki link** is a link from an article in one language edition of Wikipedia to a corresponding article (i.e., on the same subject) in another language edition.

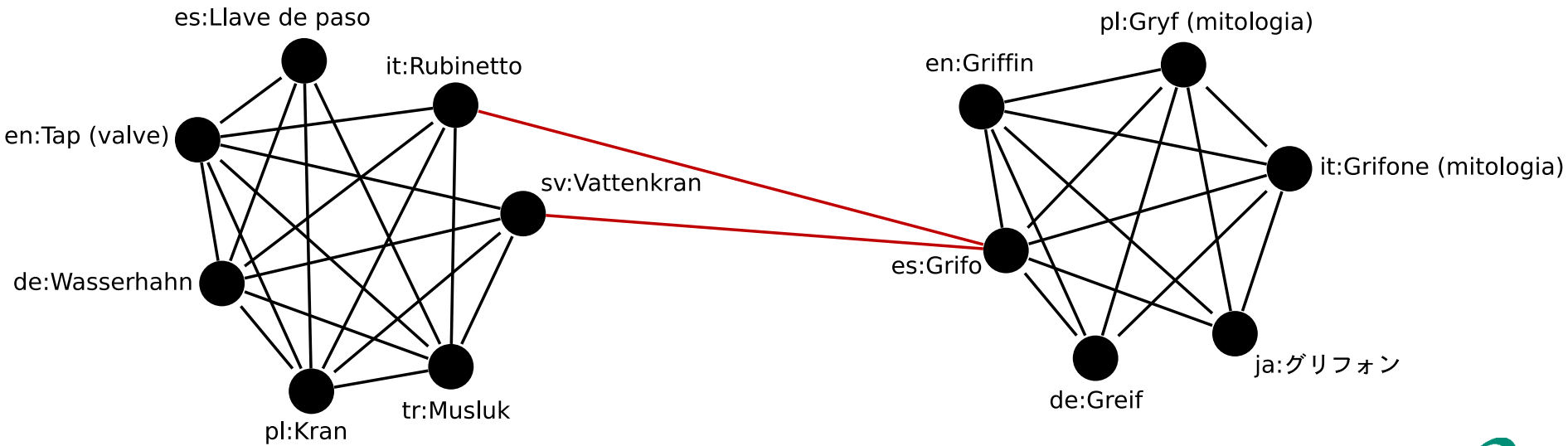
In other words, it's a kind of **translation**.

In a broader sense, interlanguage links also apply to categories, templates, etc.

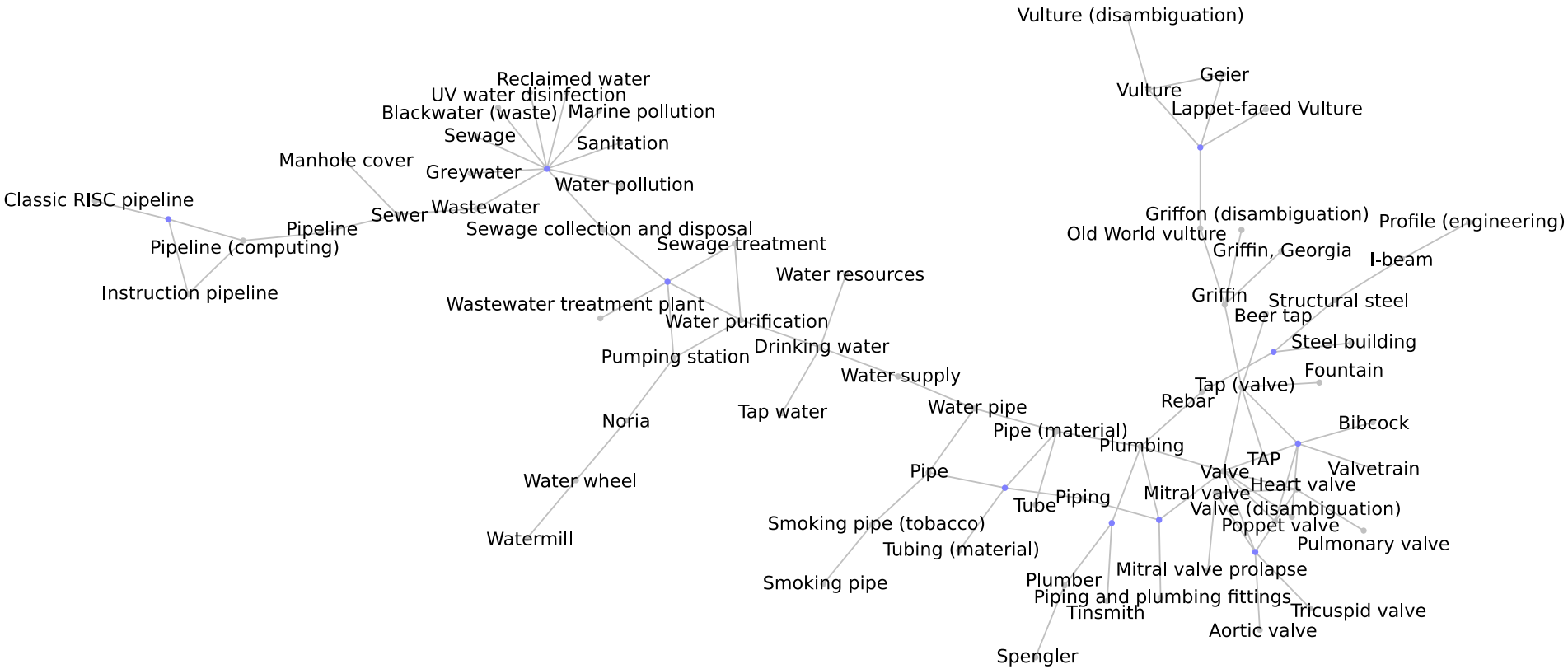
Problems with interwiki

MediaWiki engine does not guaranee the coherence of interwiki – no centralization. Instead, each autonomous language edition stores "it's own" outgoing links. Given **11.5m** articles and **90m** interwiki, the lack of centralization creates a lot of problems.

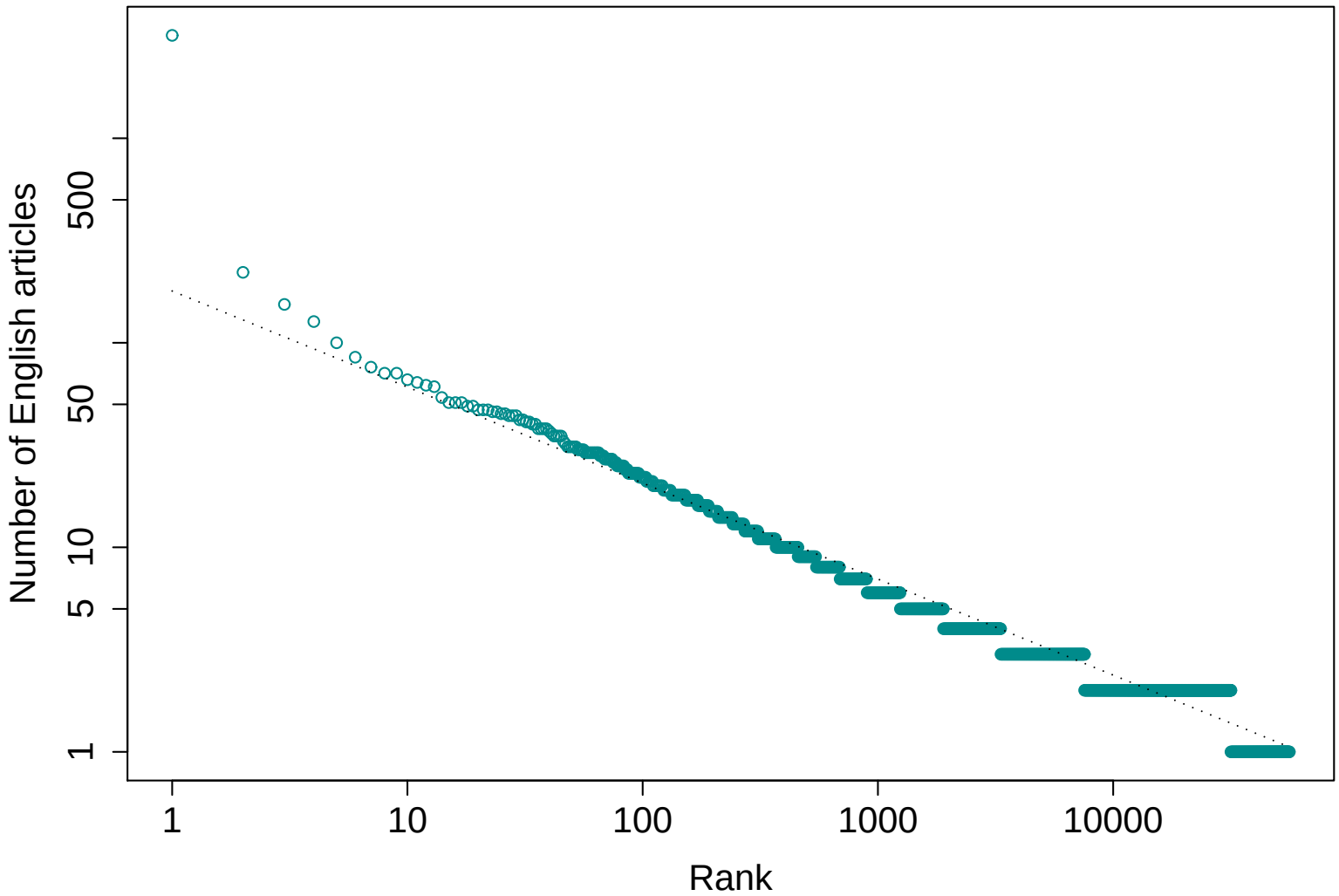
Two classes of problems: missing links and conflicting links.



Scale of the problem



Scale of the problem

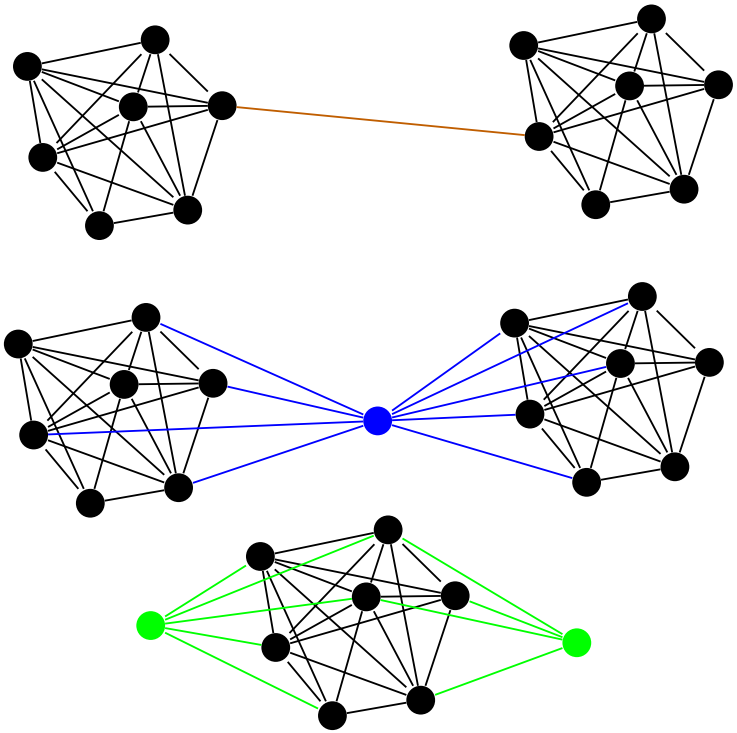


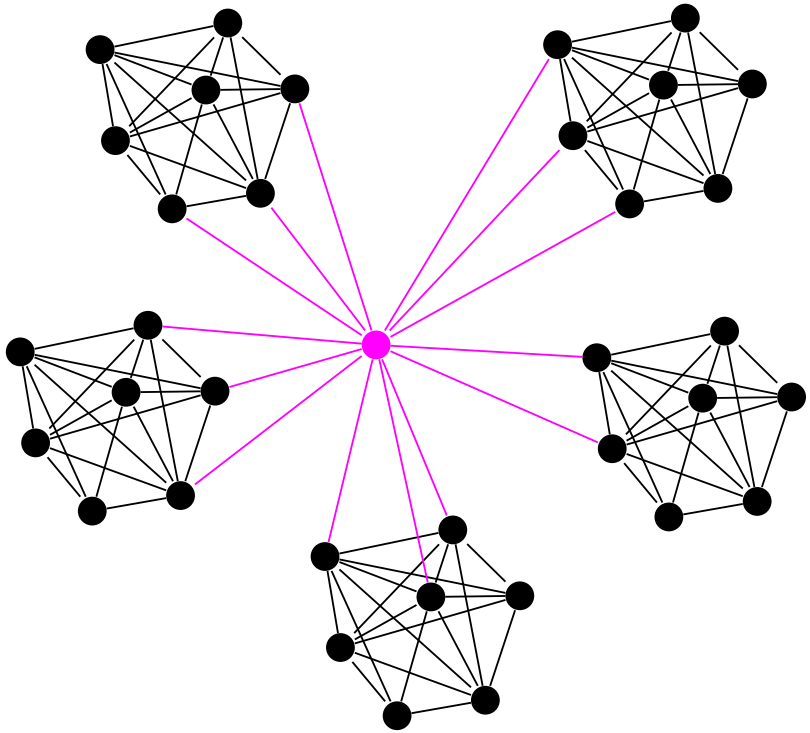
Scale of the problem

The largest component consists of approx. **80k** articles, including over **3.7k** in English (and thus at least as much different topics), and over **3.7m** interwiki links. Manual verification of all the links in no longer feasible.

Apart from that, there are over **65k** other components with conflicts.

The most common patterns





Classification of incoherent interwikis (an attempt)

- vandalism, blatant mistakes
  (ro:Nicolae Steinhardt → de:Penis; fr:Rick Ankiel → ja:日本語)
- date-related mistakes, „copy-and-paste", „off-by-one"
  (wuu:5月26号 → bn:??? ← wuu:5月27号; az:3 iyun → su:3 Juli)
  (en:10s BC ⟶ en:20s BC ⟶ en:30s BC ⟶ . . .)
- links to disambig. pages, homonyms
  (la:Benedictus (nomen) → en:Benedict; it:Rubinetto → es:Grifo)
- broader/narrower meanings, different granularity of language editions
  (pl:Województwo krakowskie (I Rzeczpospolita) → en:Kraków Voivodeship (14th century-1795) → pt:Voivodia da Cracóvia → pl:Województwo krakowskie)
- interwiki + redirect
  (en:Mother-in-law → ru:Тёща ⇝ ru:Родство → en:Kinship)
- cultural differences, limitations of the translation process
  (en:Pierogi — en:Cepelinai — en:Dumpling — en:Vareniki — en:Kalduny)

How to correct?

1. find components with conflicts
2. find a minimum set of cuts + divide into meanings
3. present the results (wikitools.icm.edu.pl)
4. **verify the proposed division into meanings**
5. **automatically remove the "inter-meaning" links introducing conflicts**

The tool

- WWW service: **http://wikitools.icm.edu.pl/**
- All the interwiki conflicts together with recommendations
  (based on dumps from November 2008)
- Directly applicable to: Wikisources, Wikibooks, Wikipedia categories, Wikia projects

The end

Thank you