

# Extracting bilingual wordlists from Wikipedia

Francis M. Tyers<sup>1,2</sup> and Jacques A. Pienaar<sup>3</sup>

<sup>1</sup>Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant,  
E-03071 Alacant (Spain)

<sup>2</sup>Prompsit Language Engineering, S.L., E-03690 St. Vicent del Raspeig (Spain)

<sup>3</sup>Centre for Text Technology, Language and literature in the South African context,  
North-West University, Potchefstroom 2520 (South Africa)

27th May 2008

# Contents

## 1 Introduction

- Bilingual wordlists
- Scarcity of resources
- Wikipedia
- Interwiki links
- Ambiguity

## 2 Algorithm

- Requirements
- Description

## 3 Method

- Method
- Results
- Analysis

## 4 Discussion

- Discussion
- Future research
- Conclusions

# What is a bilingual wordlist

group	groep
software	sagteware
programmer	programmeerder
chatroom	kletskamer
developer	ontwikkelaar
planet	planeet
solar system	sonnestelsel
god	god
universe	heelal
middle	middel
astronomer	sterrekundige
life	lewe
page	bladsy

# Bilingual wordlists

Why do we want bilingual wordlists ?

- Cross-language information retrieval (CLIR)
  - Finding search results across languages.
  - Finding aligned phrase pairs.
- Machine translation
  - RBMT: start of a bilingual part-of-speech tagged dictionary.
  - SMT: improving the quality of translation models.

## Scarcity of resources

For lesser-used languages resources (e.g. bilingual wordlists) are often scarce, they may be:

### Inexistent

- Catalan–Faroese,
- Basque–Russian etc.

### Unavailable

- not available online
- available at (great) cost, or
- in legacy encodings or formats.

### Prohibitively licensed

- non-commercial
- research and education only, or
- fully proprietary.

# What is Wikipedia? (I)

- An online, collaboratively edited encyclopaedia – content created and managed by volunteers
- Articles are available in over 250 languages – including many lesser-used and under-resourced languages
- Liberally licensed – it is freely available and freely distributable both for non-commercial and commercial use
- Requires no specific expertise other than ability to use web browser and using a simple markup language to create entries

# WIKIPEDIA

## English

*The Free Encyclopedia*  
2 284 000+ articles

## Deutsch

*Die freie Enzyklopädie*  
724 000+ Artikel

## Français

*L'encyclopédie libre*  
636 000+ articles

## Polski

*Wolna encyklopedia*  
481 000+ hasel

## 日本語

フリー百科事典  
477 000+ 記事



## Italiano

*L'enciclopedia libera*  
426 000+ voci

## Nederlands

*De vrije encyclopedie*  
417 000+ artikelen

## Português

*A enciclopédia livre*  
366 000+ artigos

## Español

*La enciclopedia libre*  
343 000+ artículos

## Svenska

*Den fria encyklopedin*  
279 000+ artiklar

search • suche • rechercher • szukaj • 検索 • ricerca • zoeken • busca  
buscar • sòk • поиск • 搜索 • søk • haku • suk • cerca • căutare • ara

<input type="text"/>	English	>
----------------------	---------	---

100 000+

Català · Deutsch · English · Español · Français · Italiano · Nederlands · 日本語 · Norsk (bokmål) · Polski · Português · Русский · Română · Suomi · Svenska · Türkçe · Volapük · 中文

10 000+

العربية · Asturianu · Kreyòl Ayisyen · Azərbaycan / آذربایجان دلی / آذربایجا · Беларуская (Акадэмічная) · বিষ্ণুপ্রিয়া মণিপুরী · Bosanski · Brezhoneg · Български · Česky · Cymraeg · Dansk · Eesti · Ελληνικά · Esperanto · Euskara · فارسی · Galego · 한국어 · हिन्दी · Hrvatski · Ido · Bahasa Indonesia · Íslenska · עברית · Basa Jawa · ქართული · Kurdi / كوردی · Latina · Lumbaart · Latviešu · Lëtzebuergesch · Lietuvių · Magyar · Македонски · मराठी · Bahasa Melayu · नेपाल भाषा · Norsk (nynorsk) · Nnapulitano · Occitan · Piemontèis · Plattdüütsch · Shqip · Sicilianu · Simple English · Sinugboanon · Slovenčina · Slovenščina · Српски · Srpskohrvatski / Српскохрватски · Basa Sunda · Tagalog · தமிழ் · తెలుగు · ไทย · Українська · Tiếng Việt

1 000+

Afrikaans · Alemannisch · Авар · Aragonés · Armãeashce · Arpitan · Bân-lâm-gú · Basa Banyumasan · Беларуская (Тарашкевіца) · भोजपुरी · Boarisch · Corsu · Чăваш · Deutsch · دۆنر · Eald Englisc · Féroyiskt · Frysk · Furlan · Gaeilge · Gàidhlig · 古文 / 文言文 · 'Ōlelo Hawai'i · Žujbntŭ · Hornjoserbsce · Ilokano · Interlingua · Ирон æвзаг · ភ្នំពេញ · Kapampangan · Kaszëbsczi · Kernewek · ភាសាខ្មែរ · Ladino · Լատին · Ligure · Limburgs · Lingála · ភាសាខ្មែរ · Malti · Māori · Монгол · Nāhuatlāhtōlli · Nedersaksisch · नेपाली · Nouormand · Novial · O'zbek · पाणि · Pangsasinān · پښتو · Қазақша · Ripoarisch · Rumantsch · Runa Simi · संस्कृत · Sámegiella · Scots · Kiswahili · Tarandine · Tatarça · Тоҷикӣ · Lea faka-Tonga · Türkmen · اردو · Veneto · Võro · Walon · West-Vlams · Winaray · 吳語 · 𑌂𑌄𑌃𑌅 · Yorùbá · Zazaki · Žemaitėška

100+

عند · Awañit'ë · Авар · Ayмара · Bamanankan · Башҡорт · Bilkol Central · 𑌂𑌄𑌃𑌅 · Chavacano de Zamboanga · Diné Bizaad · Dolnoserbski · Emiḡlān-Rumagnòl · Euegbe · Gaelg · 𑌂𑌄𑌃𑌅 · 𑌂𑌄𑌃𑌅 · 𑌂𑌄𑌃𑌅 · Hakkâ-kâ-fa / 客家話 · Igbo · 𑌂𑌄𑌃𑌅 / Inuktitut · Interlingue · 𑌂𑌄𑌃𑌅 / 𑌂𑌄𑌃𑌅 · Kongo · Кыргызча · 𑌂𑌄𑌃𑌅 · Iojban · Malagasy · Māzerūnī / مازرونی · Ming-dǝng-ngŭ · Молдовеняскэ · 𑌂𑌄𑌃𑌅 · Ekakairū Naoero · Nēhnyawēwin / 𑌂𑌄𑌃𑌅 · Norfolk / Pitkern · Нохчийн · 𑌂𑌄𑌃𑌅 · Afaan Oromoo · 𑌂𑌄𑌃𑌅 · पंजाबी / پنجابی · Papiamentu · Qimtitatarca · Romani / 𑌂𑌄𑌃𑌅 · Kinyarwanda · Gagana Sāmoa · Sardu · Seeltersk · 𑌂𑌄𑌃𑌅 · 𑌂𑌄𑌃𑌅 · Слозьньскъ · Af Soomaali · SiSwati · Reo Tahiti · Taqbaylit · Tetun · 𑌂𑌄𑌃𑌅 · Tok Pisin · 𑌂𑌄𑌃𑌅 · 𑌂𑌄𑌃𑌅 · Uyghur · Tshivenḡa · Wolof · IsiXhosa · Zeëuws · IsiZulu

# What's interesting about Wikipedia?

- Wikipedia encyclopaedias are freely-editable
- Content and structure amenable to linguistic research
- Breadth of language coverage make useful/appropriate for creating linguistic resources
- Inter-language (interwiki) page links

# What are interwiki links?

<ul style="list-style-type: none"><li>Printable version</li><li>Permanent link</li><li>Cite this page</li></ul>	<ul style="list-style-type: none"><li>5 Description and description</li><li>6 Speech versus writing</li><li>7 History of linguistics</li><li>8 See also<ul style="list-style-type: none"><li>8.1 Lists</li><li>8.2 Related topics</li></ul></li><li>9 References<ul style="list-style-type: none"><li>9.1 Textbooks</li><li>9.2 Academic works</li><li>9.3 Popular works</li><li>9.4 Reference books</li></ul></li><li>10 External links</li></ul>
---	--

languages

- af: Taalwetenskappe
- am: ብሔራዊ ጥናት
- ar: لسانيات
- an: Lingüística
- ast: Lingüística
- gn: Ñe'êkuaaty ha Ñe'êtekuaa
- bm: Kankalan
- bn: ভাষাবিজ্ঞান
- be: Мовазнаўства
- br: Yezhoniezh
- bg: Езикознание
- ca: Lingüística
- cv: Лингвистика
- ceb: Linggwistiks
- cs: Lingvistika
- co: Linguistica
- cy: Ieithyddiaeth
- da: Sprogforskning
- de: Sprachwissenschaft

## Divisions, specialties, and subfields

The field of general linguistics traditionally attempts to characterize the nature of hu it is an individual *knows* when said to know a language; and to explain how individua

All humans (setting aside extremely pathological cases) achieve competence in wh [signed languages](#)) around them when growing up, with apparently little need for cor linguists assume, the ability to acquire and use language is an innate, biologically-b ability to walk. There is no discernible *genetic* process responsible for differences be language(s) they are exposed to as a child, regardless of parentage or ethnic origin nativist schools of linguistics.

# What are interwiki links?

- Link between same article in different languages
- Links are periodically maintained by both human editors and bots<sup>1</sup>
- Harvesting these links provide useful translation equivalents

---

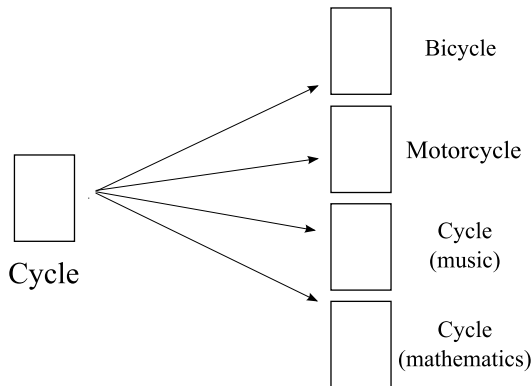
<sup>1</sup>As used by Wikipedia, a bot is a software program that makes automated changes to Wikipedia.

# Interwiki links: How links are maintained

- Manual
  - Normally done when someone starts a new article
  - Looks for the page in another language and copies in links
- Bot assisted
  - Can be supervised or unsupervised
  - Supervised mode prompts the user to resolve ambiguities
  - Unsupervised mode discards ambiguous links

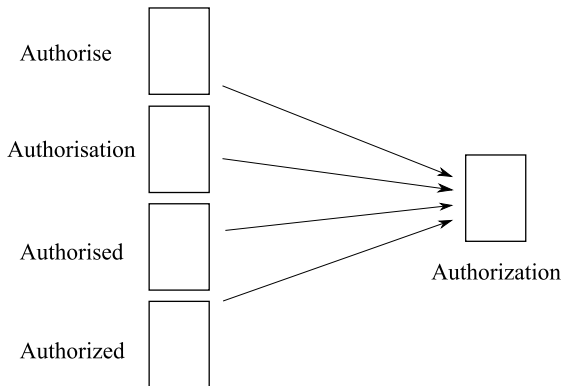
## Ambiguity: Disambiguation pages

Sometimes articles can simply be pages with lists of other articles.



## Ambiguity: Redirects

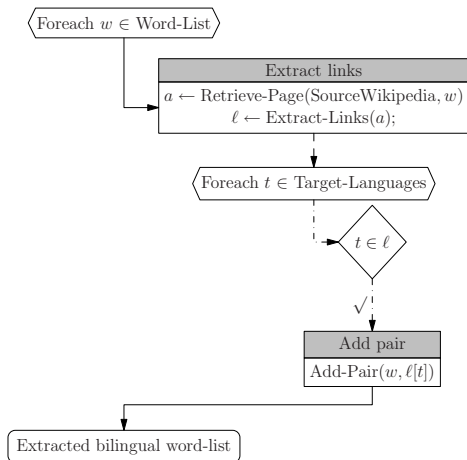
Other times, they can just *forward* the user from many article titles to one.



# Requirements

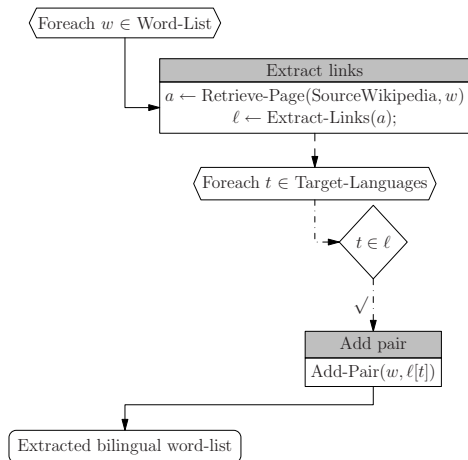
- A seed word list
  - Recommended practice: start with the better sourced language.
  - Wikipedia titles are almost exclusively nouns, therefore a list made up of nouns and proper will have greatest success.
- Computer
  - Algorithm not computationally expensive  $\implies$  desktop PC sufficient.
- Internet connection
  - Only text downloaded.
  - Low bandwidth usage and requirement.

# Algorithm



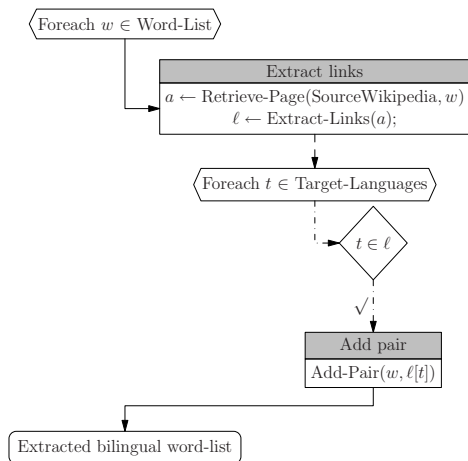
# Algorithm

cat  
computer  
house  
strip mining  
...

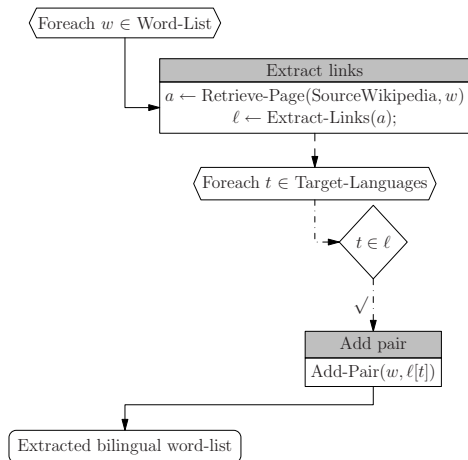


# Algorithm

cat  
computer  
house  
strip mining  
...

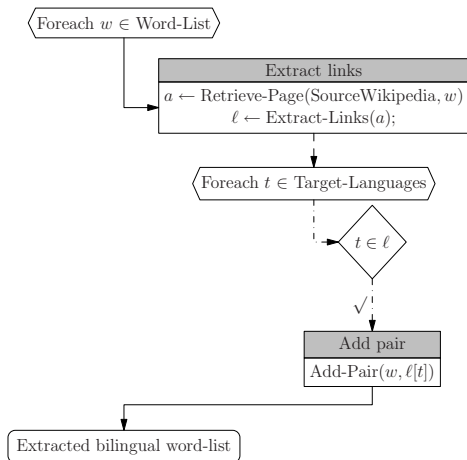


# Algorithm



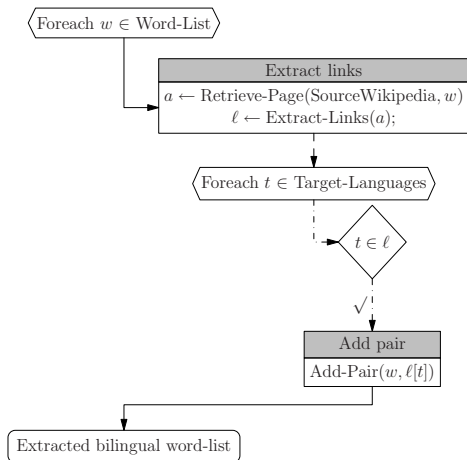
«Extract links»

# Algorithm



(ca) gat  
(fr) chat  
(af) kat  
(cy) cath

# Algorithm



cat – gat

# Method

- Wordlist of 11,393 English nouns – these were all lemmas.
- Extracted from the `apertium-en-ca` language pair – slightly biased towards scientific and technical vocabulary.
- Rationale: A post-edited list would be immediately useful in an Apertium translation pair.
- 10,024 of the 11,393 lemmas were entries in the English Wikipedia.
- Generated word lists (in Swedish, Macedonian, Afrikaans and Iranian Persian) were given to native speaker to check.
- Positive result is when translation is judged correct:
  - has the right form,
  - right sense and
  - is in correct register.

# Results

**Table:** Results for the language pairs

Language	Pages	Links	Correct	Recall	Precision
Swedish	273,291	4,913	3,428	34%	69%
Iranian Persian	32,194	1,605	1,487	14%	92%
Macedonian	14,887	779	631	6%	81%
Afrikaans	9,183	444	354	3%	79%

## Analysis of errors (I)

- (af) Right sense, wrong surface form – *vandal* translated as *vandale* (vandals).
- (sv) Right sense, wrong register – *nephrolithiasis* translated as *njursten* (kidney stone).
- (af) Wrong sense, right domain – *sociolinguist* translated as *sosiolinguistiek* (sociolinguistics).
- (af) Wrong sense, wrong domain – *solidarity* translated as *Solidarność* (a Polish trade union)

## Analysis of errors (II)

- Errors generally found to exist at approximately same frequency.
- None particularly more frequent than others.
- The errors examined seem to have been caused by redirects.
- *No full quantitative analysis done.*

# Discussion

- Method provides basic bilingual word list.
- Precision is good but recall low.
- Recall will increase over time as more articles are added, and the link coverage improves.

## Further and future research

- Double-check each pair – Ensuring that a retrieved link points back to the same source.
- Avoid following redirects.
- Retrieve gender information from article leads.
- Use category of Wikipedia pages.
- Creation of gazeteers.
- More thorough evaluation.

# Conclusions

- Presented simple, computationally inexpensive and fast means of automatically obtaining bilingual word list.
- Good accuracy (lowest 69%) obtained.
- Poor recall.
- Method could be useful for bootstrapping more complex induction techniques for under-resourced languages.