# Automatic Methods for the Extension of a Bilingual Dictionary Using Comparable Corpora

Michael Rosner & Kurt Sultana
University of Malta

## Introduction

Machine-readable dictionaries are of fundamental importance for NLP since these are required for applications such as word sense disambiguation, query expansion and text classification amongst others.

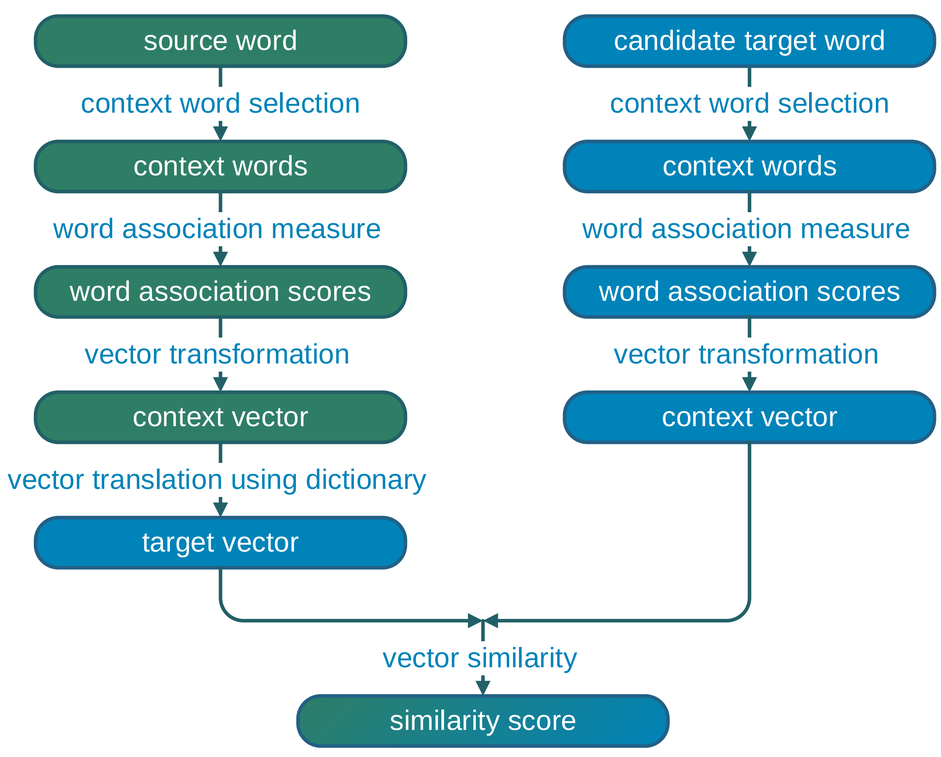
No bilingual dictionary is complete since languages are in a constant state of change. Additionally, dictionaries are unlikely to achieve complete coverage of all language terms. For these reasons, methods for reliable dictionary expansion are crucial. However, manual dictionary extension is laborious, expensive and highly time-consuming. Automatic methods for dictionary extension are highly desirable.

Word alignment techniques are available for extracting translation pairs from parallel corpora. However, parallel corpora are difficult to obtain or create. For this reason, we require methods for the extraction of translation pairs from non-aligned corpora that are easier to obtain. Such methods are based on the hypothesis that words with same meaning share similar contexts.

### Scope

Our research concerns the investigation and development of a framework for extending computational dictionaries efficiently and reliably, making use of available resources in the best way possible.

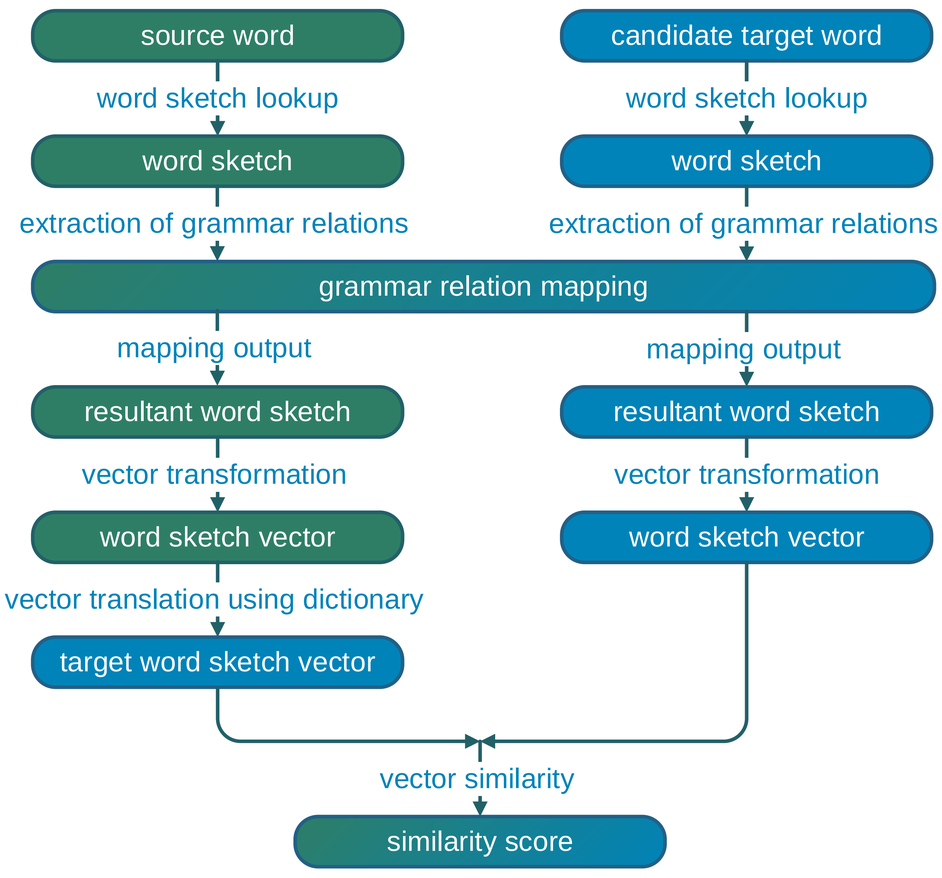
## Method 1 – Context Vector Method




## Method 2 – Word Sketch Method

### Word Sketch

A word sketch is a "one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour" (Rychlý and Kilgarriff, 2007).

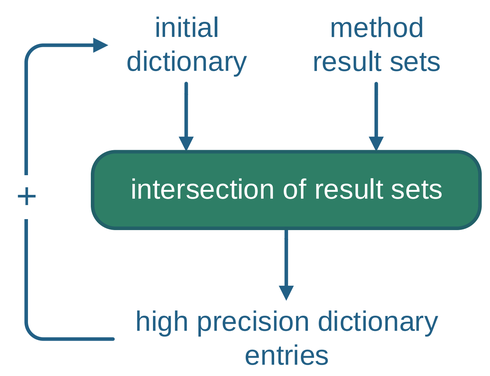

## High-precision dictionary entries

Dictionary entries added to the current dictionary must be as reliable and accurate as possible.

We achieve high precision by:

(i) intersecting results sets from methods to obtain the resultant dictionary entries

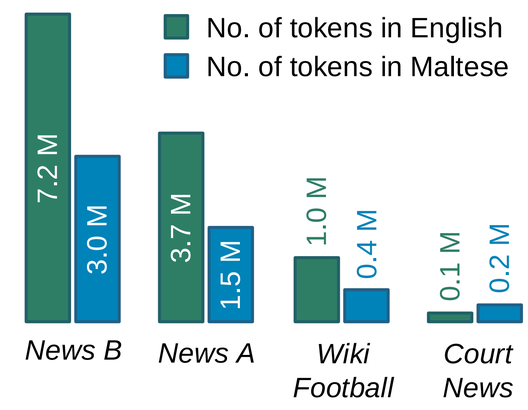
(ii) iterating the extraction process




## Data Sets

The Falzon dictionary (Falzon, 1987) was used as the initial dictionary. It contains approximately 5,400 entries. Its limited size made it an excellent candidate for the research carried out.

News text and *Wikipedia* were used as the main sources for comparable text. Domain-specialised corpora (sports, court news) were also used for evaluation.




## Results

### Context Vector Method

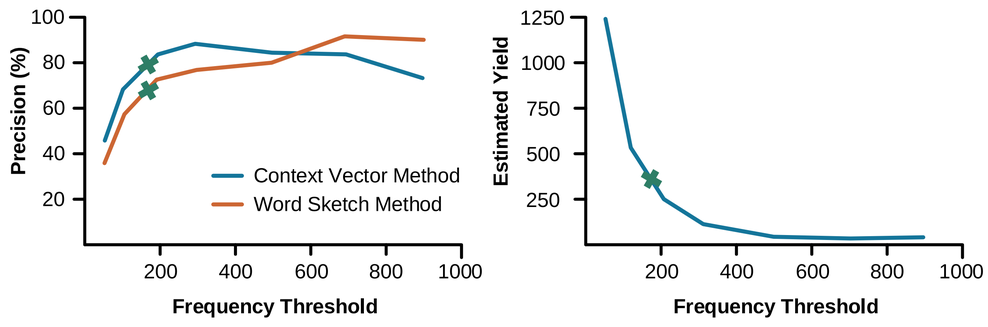
- maximum precision of **78.8%**
- frequency threshold of **150**, first-time yield of **180** word pairs
- performs well even with **smaller and specialised** corpora
- optimum context window size of **3 to 4 terms** i.e. context is taken as 3 to 4 terms to the left and to the right of the source/target word

### Word Sketch Method

- maximum precision of **73.2%**
- frequency threshold of **150**, first-time yield of **180** word pairs
- does not perform so well with smaller corpora – word sketches are sensitive to corpus size
- no context window required – word sketches pick adjacent terms according to grammar relations

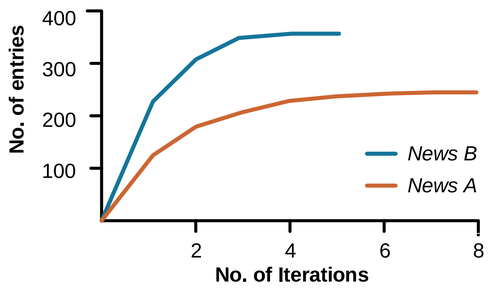
### Effect of Word Frequency

The more frequent a word is, the more context it would have. It was confirmed that if the frequency threshold used for choosing source words is increased, precision increases but yield decreases. However, using method combination with iterations, yield can be increased.




### High-precision dictionary entries using method combination with iterations

- maximum precision of **81.7%**
- frequency threshold as low as **20**
- maximum output of **344** resultant dictionary entries
- other clues such as **alignments** or **cognates** could be used




### Key points for dictionary extraction from comparable corpora

- Does corpus contain an adequate amount of **parallel text**? Consider alignment?
- What is the **size** of the corpus?
- What **resources** are available for the languages? Sketch grammars, lemmatisers, POS-taggers?
- Is the corpus **specialised** to a particular domain?
- What **precision** is expected? Method combination gives higher precision.
- What **yield** is expected? Single methods give higher yield.

**References**: Pavel Rychlý and Adam Kilgarriff. 2007. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 41–44, Prague, Czech Republic, June. Association for Computational Linguistics