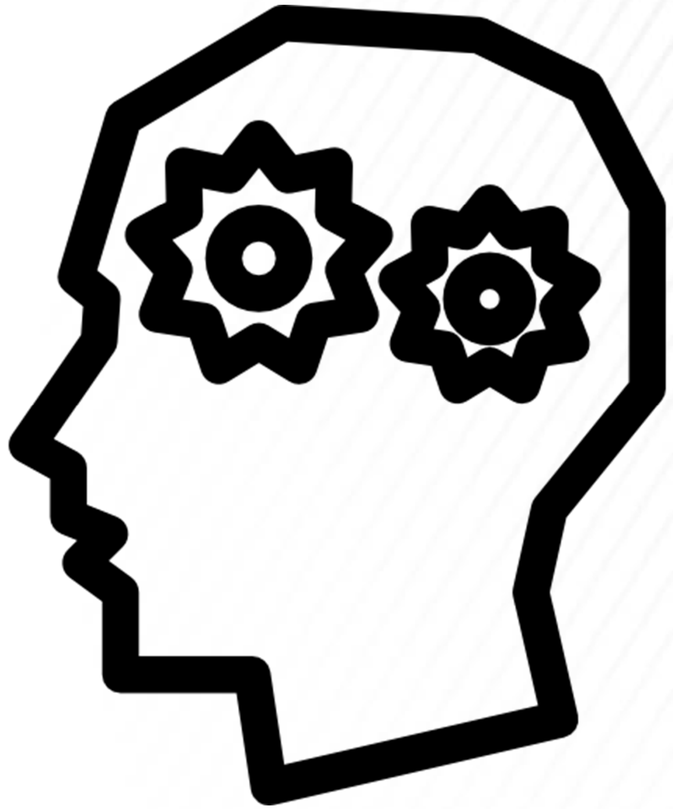
# Automated translation analysis using a multi-faceted framework



Rafał Jaworski

Andrzej Zydroń

XTM International

# Word Similarities

## Word-level alignment

- Word-level alignment is a process of automatic matching of similar words.
- For a pair of sentences the algorithm must decide which words from the source and target sentences are each other’s counterparts:

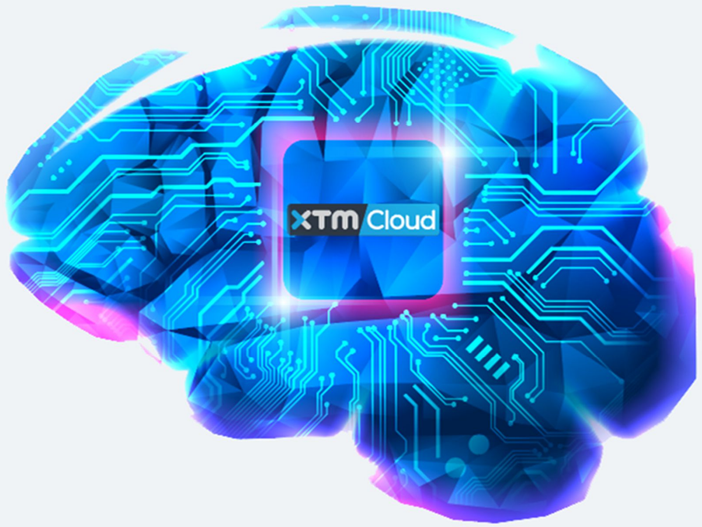
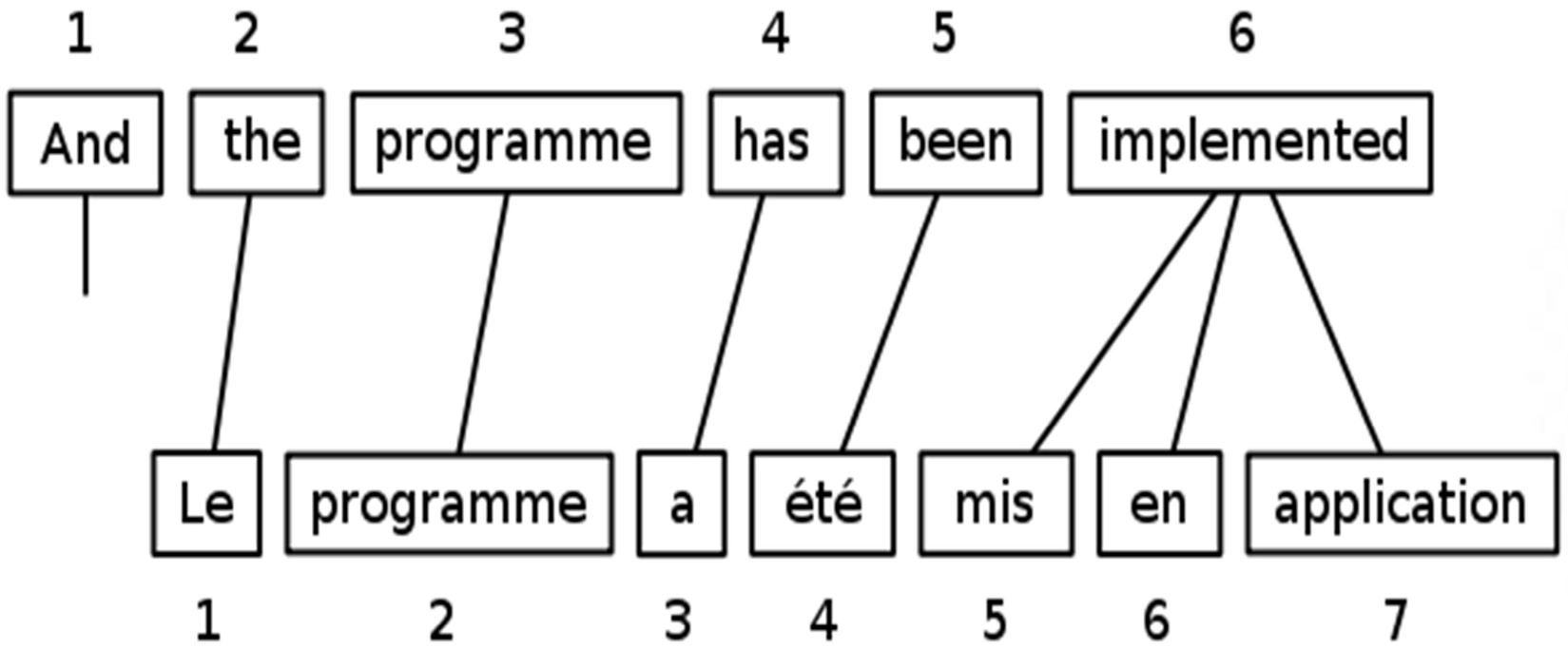



*Image source: wikipedia.org*

# Word Similarities

## Word-level alignment



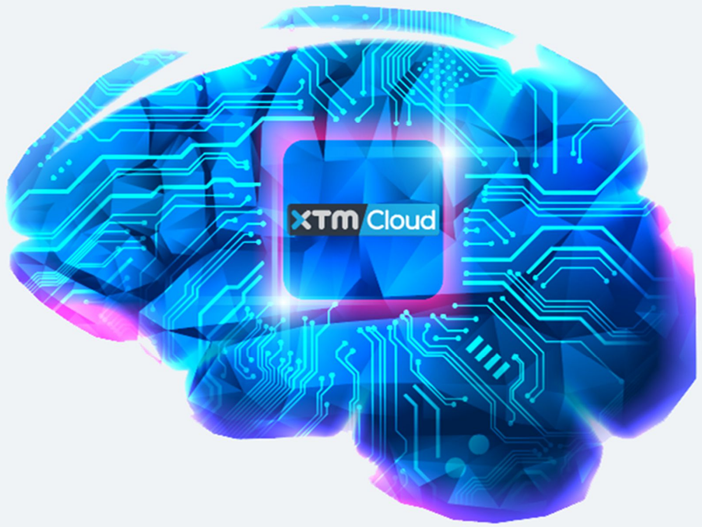
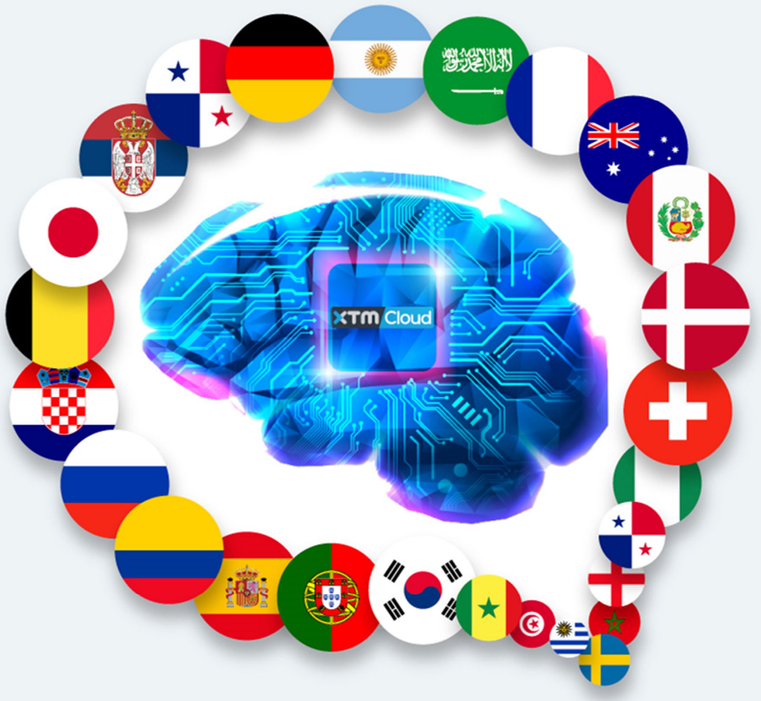




*Image source: wikipedia.org*

# Word Similarities

## Vector Space

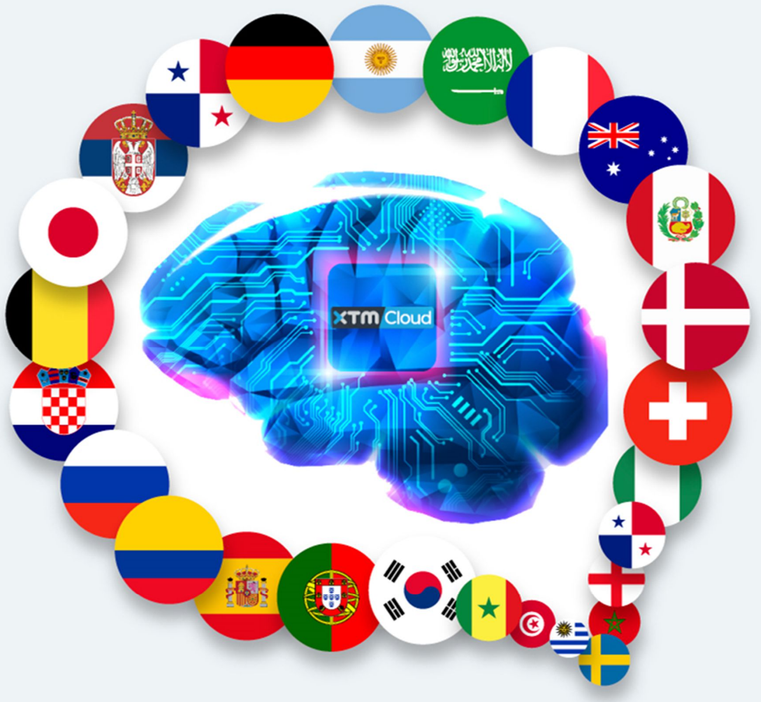
- In 2016 Facebook Research published further work on Vector Space
- Based on a crawl of the entire Internet
- Vector Space for **157** languages



# Word Similarities

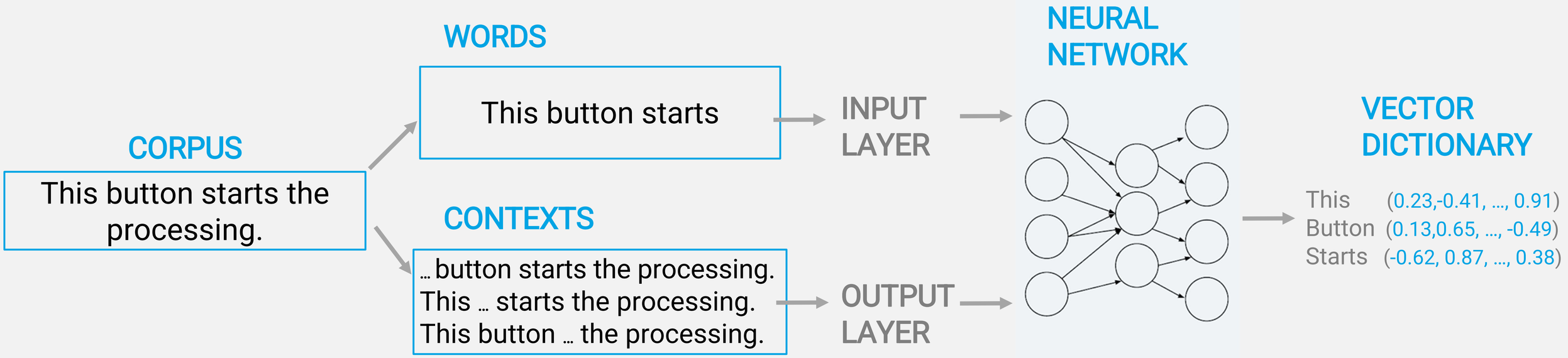
## Vector Space

- The most complete **Vector Space language model for each of the 157 languages**
- Individual language Vector Spaces are unique: not directly comparable
- In 2017 Babylon Health produced a paper showing a possible way of ‘normalizing’ the Vector Space between 2 languages



# Word Similarities

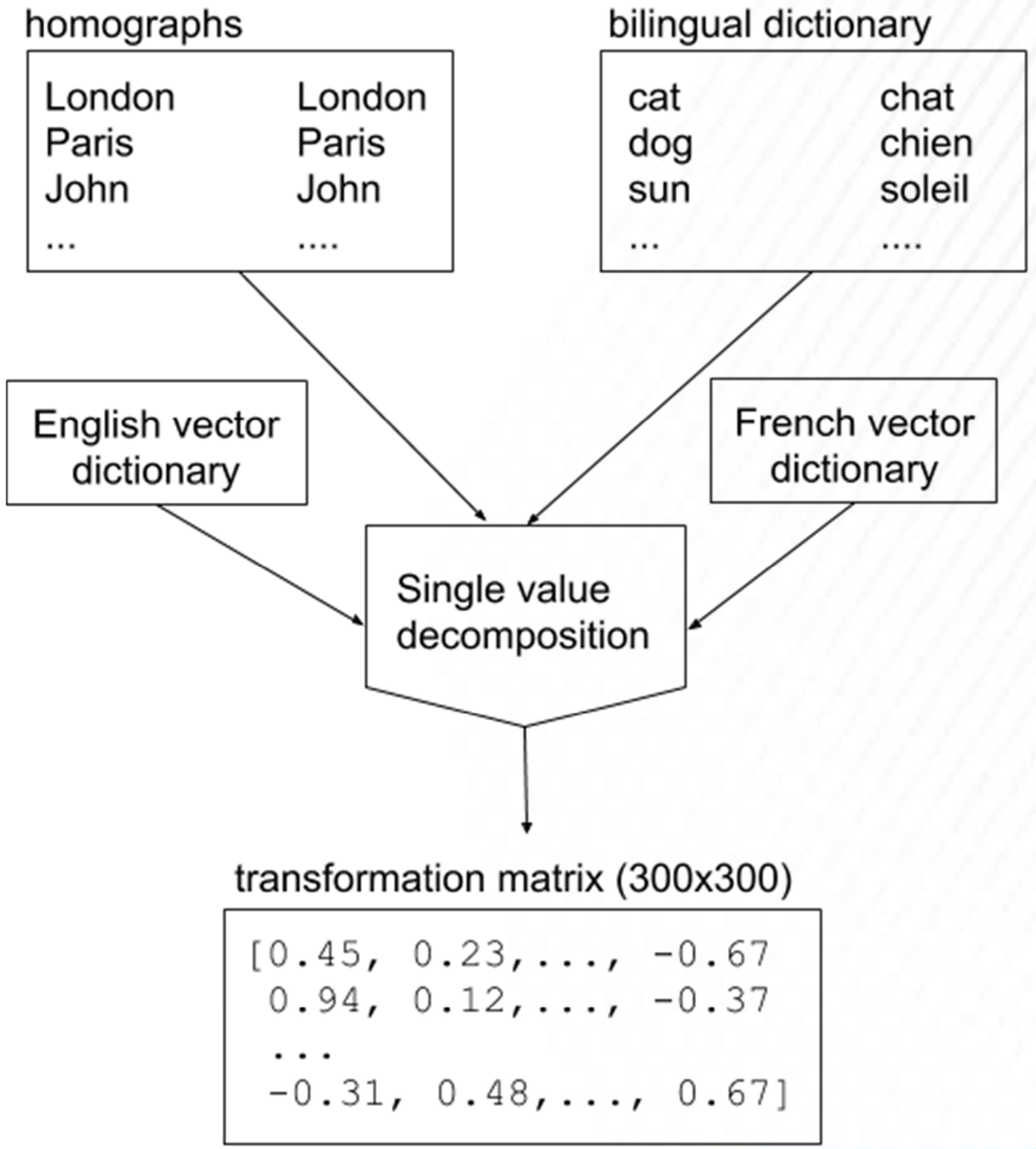
## Vector Space



# Word Similarities

## Inter-language Vector Space

### Transformation Matrix

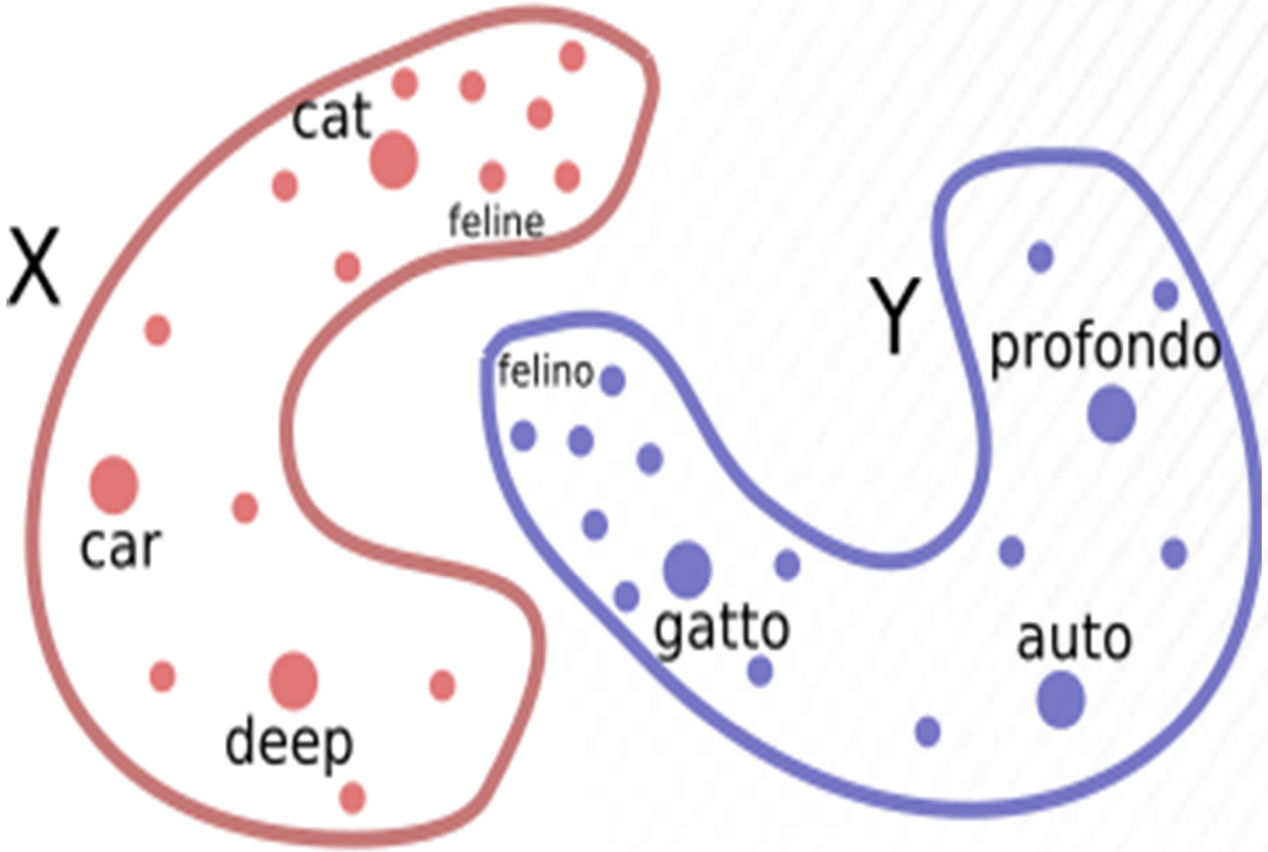



XTM International

# Word Similarities

## Inter-language Vector Space

Transformation Matrix



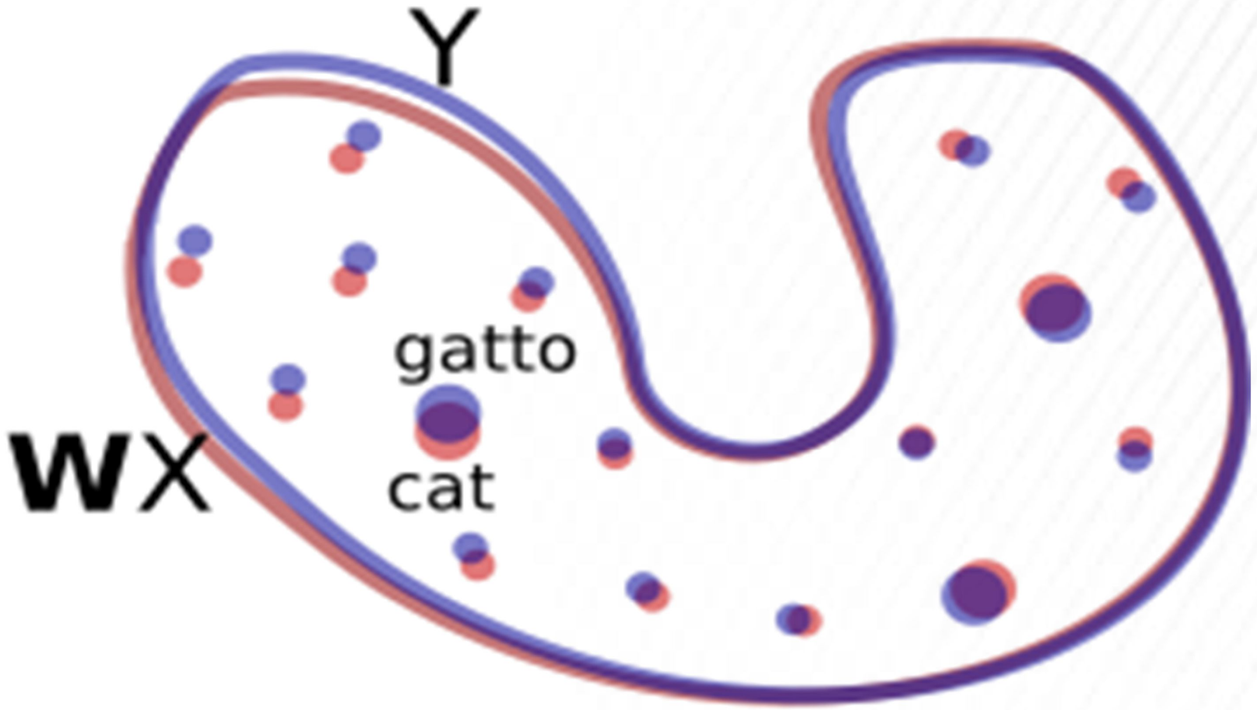

XTM International

*Image source: Facebook MUSE*

# Word Similarities

## Inter-language Vector Space

Transformation Matrix



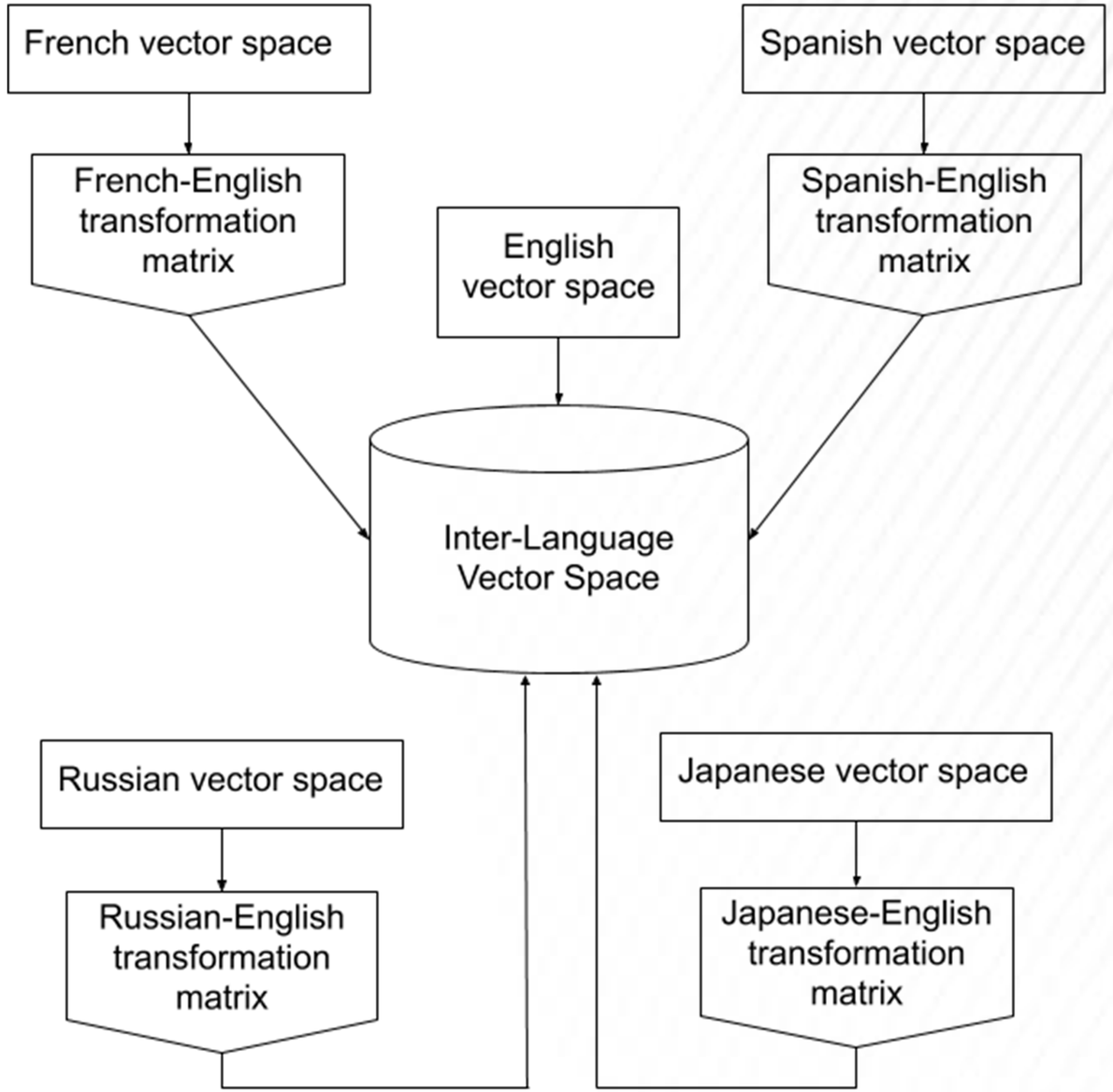

XTM International

*Image source: Facebook MUSE*

# Word Similarities

## Inter-language Vector Space

### Transformation Matrix




XTM International

# Word Similarities

## Inter-language Vector Space

- **Unique XTM development**/contribution
- Normalized Vector Space for **50 languages** onto the same plane
- Uses the XTM bilingual dictionaries generated from BabelNet and other online resources
- **US Patent** applied for
- Computes **similarity score** for each source-target word pair

Powered by XTM AI


XTM International

# Word Similarities

## Inter-language Vector Space

XTM International

| English word | Italian word | Similarity |
|---|---|---|
| cat | gatto | 0.696 |
| cat | gatta (female cat) | 0.552 |
| cat | giorno (day) | 0.164 (as expected, low similarity) |
| day | giorno | 0.692 |
| day | fuoco (fire) | 0.193 (as expected, low similarity) |
| fire | fuoco | 0.590 |

# Word Similarities

## Inter-language Vector Space

XTM International

Matching words in Marathi

| | मागील | महिन्यात | खेळलेल्या | खेळाची | आकडेवारी |
|---|---|---|---|---|---|
| Statistics | | | | | |
| of | | | | | |
| the | | | | | |
| games | | | | | |
| played | | | | | |
| during | | | | | |
| last | | | | | |
| month | | | | | |

# Word Similarities

## Inter-language Vector Space

XTM International

Matching words in Kannada

| | ಸ್ಥಿತಿ | ಪಟ್ಟಿಯನ್ನು | ತೋರಿಸಬೇಕೆ | ಅಥವ | ಬೇಡವೆ |
|---|---|---|---|---|---|
| Whether | | | | | ■ |
| or | | | | ■ | |
| not | | | | | |
| to | | | | | |
| show | | | | | |
| the | | ■ | | | |
| status | ■ | | | | |
| bar | | | | | |

# Multi-faceted Translation Analysis

## MFTA

| | | ils | connaissent | très | bien | le | problème de surproduction | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | problème | de | surproduction | . |
| | They | 0.88 ✚ | 0.48 | 0.46 ⇳ | 0.56 ⇳ | 0.2 | 0.22 ↔ | 0.3 | 0.15 | 0.16 | |
| know about | | 0.37 ↕ | 0.48 ✢ | 0.25 ↕ | 0.43 ↕ | 0.19 ↕ | 0.27 ↔ | 0.33 | 0.16 ↕ | 0.08 | |
| know about | know | 0.37 | 0.68 ✚ | 0.31 | 0.49 | 0.19 | 0.27 ↔ | 0.33 | 0.16 | 0.08 | |
| know about | about | 0.24 | 0.36 | 0.23 | 0.39 | 0.16 | 0.19 ↔ | 0.25 | 0.41 ⇔ | 0.04 | |
| the overproduction | | 0.12 | 0.16 | 0.15 | 0.16 | 0.14 | 0.6 ↔ | 0.31 | 0.11 | 0.66 ✢ | |
| the overproduction | the | 0.23 | 0.35 | 0.23 | 0.34 | 0.75 ✚ | 0.36 ↔ | 0.35 | 0.44 ⇳ | 0.21 | |
| the overproduction | overproduction problem | 0.18 | 0.29 | 0.22 | 0.24 | 0.14 | 0.8 ✚ | 0.68 ↕ | 0.15 | 0.58 | |
| the overproduction | overproduction | 0.17 | 0.15 | 0.13 | 0.13 | 0.11 | 0.63 ↔ | 0.54 | 0.14 | 1.77 ✚ | |
| the overproduction | problem | 0.23 | 0.29 | 0.25 | 0.28 | 0.1 | 0.64 ↔ | 1.94 ✚ | 0.14 | 0.56 | |
| | . | | | | | | | | | | 1.46 ✚ |

# Multi-faceted Translation Analysis

## MFTA

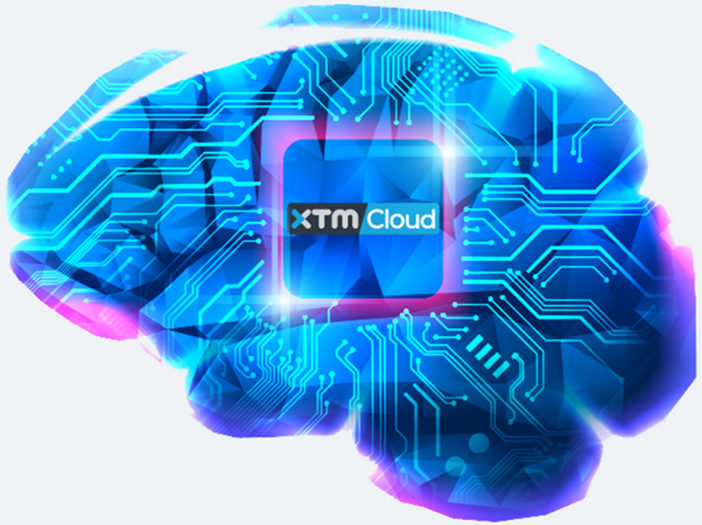
| | | Wireless | Workbench | 6 | 에는 | 주파수 | 조정 | 의 | 모든 측면 | 모든 | 측면 | 을 | 관리할 | 수 | 있는 | 강력한 | 도구 | 세트 | 가 | 포함되 | 어 | 있 | 습니다 | . |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wireless Workbench | | | | | | | | | 0.15 | | | | | | | | | | | | | | | |
| | Wireless | 1.81 I | 0.17 | | 0.12 | 0.11 | 0.09 | 0.08 | | 0.07 | 0.06 | 0.05 | 0.04 | 0.04 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | |
| | Workbench | 0.16 | 1.81 I | | 0.15 | 0.13 | 0.11 | 0.1 | | 0.08 | 0.07 | 0.06 | 0.05 | 0.04 | 0.04 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | |
| | 6 | | | 2.43 I | | | | | | | | | | | | | | | | | | | | |
| | contains | 0.11 | 0.13 | | 0.05 | 0.3 | -0.05 | 0.4 | | 0.05 | 0.3 | -0.03 | 0.3 | 0.04 | 0.08 | 0.14 | 0.3 | 0.3 | 0.39 | 0.16 IV | 0.32 | 0.08 | 0.03 | |
| | a | 0.09 | 0.1 | | 0.3 | 0.3 | 0.29 | 0.69 III | | 0.24 | 0.3 | 0.66 | 0.3 | 0.13 | 0.14 | 0.08 | 0.3 | 0.3 | 0.54 | 0.03 | 0.48 | 0.09 | 0.42 | |
| powerful set | | | | | | | | | 0.16 | | | | | | | | | | | | | | | |
| | powerful | 0.07 | 0.08 | | 0.18 | 0.3 | 0.18 | 0.2 | | 0.2 | 0.3 | 0.14 | 0.3 | 0.1 | 0.06 | 0.3 III | 0.3 | 0.3 | 0.01 | -0.06 | -0.06 | 0 | 0.09 | |
| set of tools | | | | | | | | | 0.16 | | | | | | | | | | | | | | | |
| | set | 0.06 | 0.07 | | 0.3 | 0.3 | 0.3 | 0.3 | | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.5 II | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | |
| | of | 0.05 | 0.06 | | 0.18 IV | 0.3 | 0.12 | 0.65 | | 0.18 ⇳ | 0.3 | 0.01 | 0.3 | 0.09 ⇳ | 0.11 ⇳ | 0.24 | 0.3 | 0.3 | 0.54 | 0.11 | 0.55 | 0.11 ⇳ | 0.02 | |
| | tools | 0.04 | 0.05 | | 0.3 | 0.3 | 0.3 | 0.3 | | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.54 II | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | |
| | to | 0.03 | 0.04 | | 0.19 | 0.3 | 0.14 | 0.59 | | 0.13 | 0.3 | 0.65 | 0.3 | 0.28 | 0.34 | 0.23 | 0.3 | 0.3 | 0.72 III | 0.14 | 0.58 | 0.21 | 0.56 | |
| | manage | 0.03 | 0.03 | | 0.3 | 0.3 | 0.3 | 0.3 | | 0.3 | 0.3 | 0.3 | 0.5 II | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | |
| every aspect | | | | | | | | | 0.21 III | | | | | | | | | | | | | | | |
| | every | 0.02 | 0.02 | | 0.07 | 0.3 | 0.02 | 0.48 | | 0.15 ⇔ | 0.3 | 0.02 | 0.3 | 0.04 | -0.01 | 0.25 | 0.3 | 0.3 | 0.61 | -0.02 | 0.56 | -0.01 | 0.1 ⇳ | |
| | aspect | 0.02 | 0.02 | | 0.3 | 0.3 | 0.3 | 0.3 | | 0.3 | 0.5 II | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | |
| | of | 0.01 | 0.02 | | 0.21 | 0.15 | 0.15 | 0.56 | | 0.22 | 0.15 | 0.47 | 0.15 | 0.17 | 0.2 | 0.24 | 0.15 | 0.15 | 0.65 | 0.31 | 0.67 III | 0.25 | 0.54 | |
| frequency coordination | | | | | | | | | 0.09 | | | | | | | | | | | | | | | |
| | frequency | 0.01 | 0.01 | | 0.3 | 0.54 II | 0.3 | 0.3 | | 0.3 | 0.3 | 0.3 | 0.3 | 0.5 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | |
| | coordination | 0.01 | 0.01 | | 0.09 | 0.3 | 0.26 IV | 0.11 | | 0.03 | 0.3 | 0.06 ⇳ | 0.3 | 0.06 | 0.06 | 0.12 | 0.3 | 0.3 | 0.11 | 0.02 | 0.22 | 0.06 | -0.05 | |
| | . | | | | | | | | | | | | | | | | | | | | | | | 2.44 I |

# Word Similarities

## Inter-language Vector Space / MFTA

### Recap:

- Advanced new direction for **linguistic AI**
- Based on **neural network** analysis of vast amounts of textual data: comprehensive language models
- XTM uses a crawl of the **whole Internet** for each language – terabytes of data




XTM International

# What we have now, and what's to come

## Current XTM functionalities driven by AI

- Automatic placement of inline elements
- Aligner
- Terminology extractor / bilingual terminology extractor
- Systran AI-enhanced TM
- Unique fuzzy matching algorithm

# What we have now, and what's to come

## Automatic placement of inline elements

A useful improvement of the work of a translator is the feature of automatic placement of inline elements.

Inline elements are all technical symbols that occur in the translated text such as:

- HTML/XML tags
- Rich text formatting tags
- Special symbols

# What we have now, and what's to come

## Automatic placement of inline elements

Authentication is required to  change  your own user data

Vous devez vous authentifier pour  modifier  vos propres données utilisateur

# What we have now, and what's to come

## Automatic placement of inline elements

Real-life tests over the course of two months:

- 1.3 million uses of the auto-inlines feature
- 98% accuracy
- an estimated 416 person days saved

# Future plans

## XTM AI

New functionalities will include:

- Automatic translation review
- Sub-segment matching
- Predictive typing
- Enhanced translation memory fuzzy matching

... and many more

# Q&A