# Multilingual Compound Splitting combining Language Dependent and Independent Features

Elizaveta Loginova-Clouet,
Béatrice Daille
{elizaveta.loginova, beatrice.daille}@univ-nantes.fr

University of Nantes

Dialogue 2013

# Context

TTC Project (Terminology Extraction, Translation Tools and Comparable Corpora)
http://www.ttc-project.eu

- Terminology extraction
- Alignment (translation) of terms between two languages
- Comparable corpora
- 7 languages: English, French, German, Spanish, Russian, Latvian, Chinese

# Compounds: challenge for NLP and multilingual applications

German, Dutch, Greek, Swedish, Danish, Finnish, Russian, etc.

- Machine translation: Compound term $\Leftrightarrow$ MWT
  DE "Rotorblatt" - EN "rotor blade" - FR "pale de rotor" - RU "лопасть ротора"
  DE "Windenergieanlage" - EN "wind generator" - FR "générateur éolien" - RU "ветрогенератор"
- Information retrieval
- Speech recognition
- etc.

## Difficulties

- most of compounds are not listed in lexical sources
- low frequency $\Rightarrow$ need very large training data to be observed

# Compounding

method of word formation consisting in a combination of two (or more) *autonomous* lexical elements that form a unit of meaning

- Concatenation: EN kilowatthour, FR poisson-scie [saw fish]
- Interfix
  RU "o", "e": капиталовложение [capital investment]
  kapitalovlozhenije = kapital [capital] + o + vlozhenije [investment]
  DE "s": Staatsfeind [public enemy] = Staat [state] + s + Feind [enemy]
- Stem transformations
  DE Muse**en**verwaltung [museum administration] = Muse**um** [museum] + Verwaltung [administration]
  RU ветрогенератор [wind generator]
  vet**ro**generator = vet**er** [wind] + generator [generator]
- "Neoclassical compounds" [Namer, 2009]
  EN **multi**media, DE **Turbo**maschine

# Compound splitting methods

## Methods using language-specific rules

Transformation rules [Langer, 1998] DE
"s" $\rightarrow$ "", "en" $\rightarrow$ "", "en" $\rightarrow$ "um", etc.

- Dictionary (BananaSplit, [Ott, 2005])
- Monolingual corpus (IMS Splitter, [Weller and Heid, 2012])
- Parallel corpus [Koehn and Knight, 2003]

## Language independent methods

- Automatic acquisition of transformation rules [Macherey et al., 2011]
  - Parallel corpus for training
  - Word frequency in general language
- Finding component boundaries [Hewlett et Cohen, 2011]
  - Probability of character sequences in the languages
  - Entropy as measure of probability

# Multilingual compound splitting method

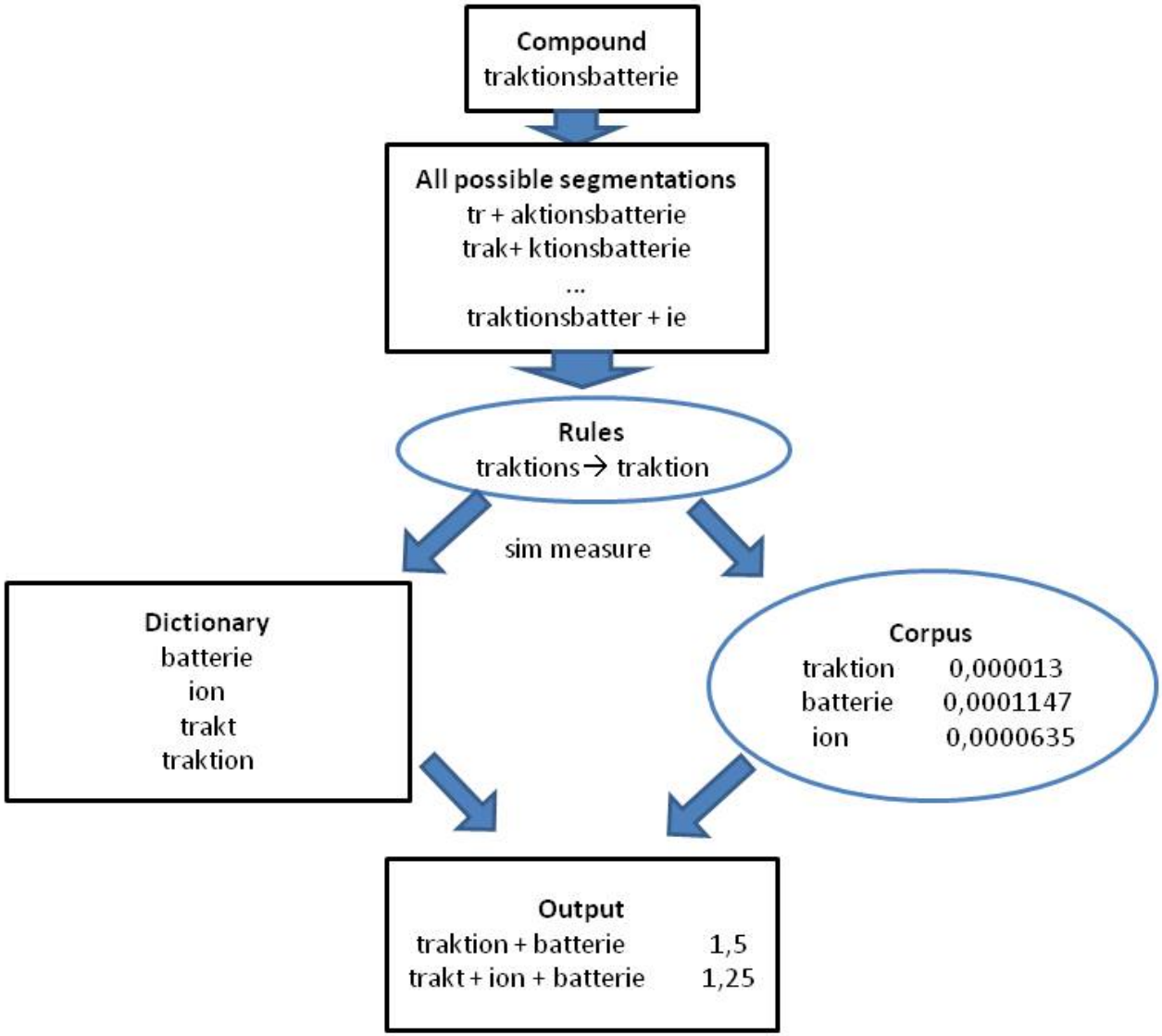
**Language independent features**

- monolingual corpus data
- similarity measure between component and its lemma

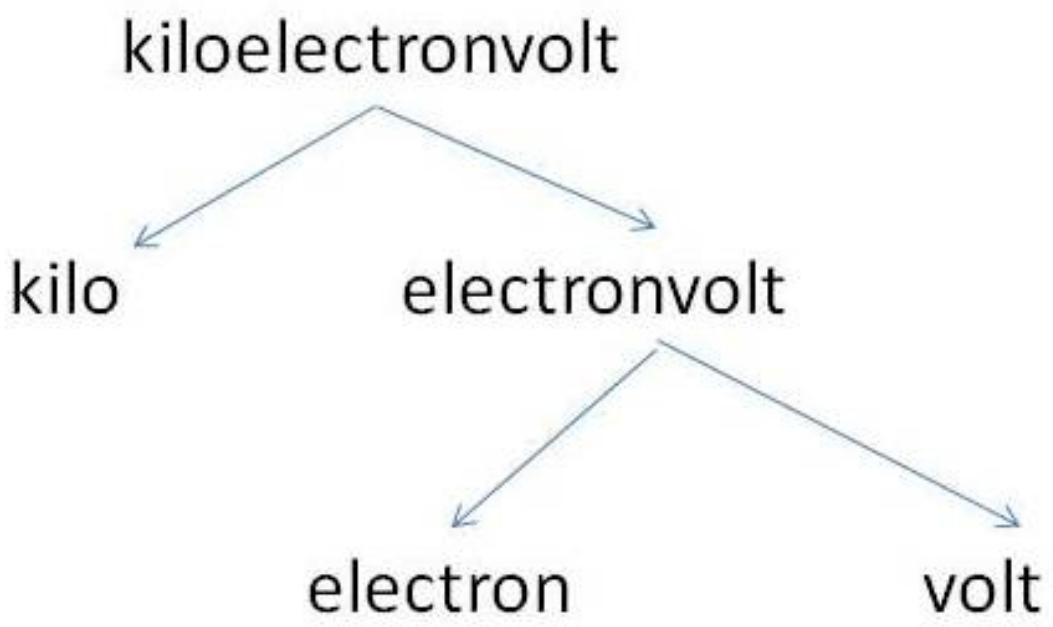
**Language-specific transformation rules** (if available)

www.lina.univ-nantes.fr/?Compound-Splitting-Tool.html

# Splitting algorithm



# Recursive splitting



# Similarity measures

## Find lemmas of components

RU vetro- → vetr → veter [wind]
DE Prozess → Prozeß

- Normalized edit distance (Levenshtein distance)

$$sim(X, Y)_{Levenshtein} = 1 - \frac{nbEditOper}{max(length(X), length(Y))}$$

- Longest common prefix

$$sim(X, Y)_{Prefix} = \frac{length(prefix)}{max(length(X), length(Y))}$$

# Experiments

**Languages:** German and Russian

## German and Russian

- Baseline (Dictionary)
- Dictionary + Similarity + Transformation rules
- Dictionary + Similarity + Transformation rules + Corpus

## Russian

- Levenshtein / Longest common prefix similarity measure
- Small rules-set / large rules-set

**Evaluation:** Accuracy in Top 1 and Top 5

$$Accuracy_{TopN} = \frac{\text{nb compounds with a correct segmentation in Top N}}{\text{total nb compounds}}$$

# Dataset

- **Compounds related to wind energy field**
  DE 446 compounds ⇐ [Weller and Heid, 2012]
  RU 348 compounds ⇐ wind energy corpus
  (among 7000 most frequent lexemes 348 are compounds)
- **Dictionaries**
  DE German part of public DE-EN dictionary Dict.cc
  RU electronic version of Ozhegov dictionary
  + list of neoclassical elements [Béchade, 1992]
- **Specialized corpora (wind energy)**
  DE 1 500 000 tokens
  RU 300 000 tokens

# German transformation rules [Langer, 1998]

| N | Transformation | Example |
|---|---|---|
| 1 | "s" → "" | Staatsfeind |
| 2 | "n" → "" | Soziologenkongreß |
| 3 | "en" → "" | Straußenei |
| 4 | "er" → "" | Geisterstunde |
| 5 | "es" → "" | Geisteshaltung |
| 6 | "nen" → "" | Wöchnerinnenheim |
| 7 | "en" → "us" | Aphorismenshatz |
| 8 | "en" → "um" | Museenverwaltung |
| 9 | "a" → "um" | Aphrodisiakaverkäufer |
| 10 | "en" → "a" | Madonnenkult |
| 11 | "e" → "" | Hundehalter |
| 12 | "en" → "on" | Stadienverbot |
| 13 | "a" → "on" | Pharmakaanalyse |
| 14 | "ien" → "" | Prinzipienreiter |
| 15 | "i" → "e" | Carabinierischule |
| 16 | "" → "en" | Südwind |
| 17 | "" → "e" | Kirchhof |

# Russian transformation rules

| N | Left context | Transformation | Example |
|---|---|---|---|
| **Small rules-set** | | | |
| 1 | - | "о" → "" | капитало- → капитал |
| 2 | - | "е" → "" | средне- → средний |
| **Large rules-set** | | | |
| 3 | - | "о" → "а" | водо- → вода |
| 4 | - | "е" → "я" | земле- → земля |
| 5 | "ж/ш/щ/ч/ц" | "е" → "а" | тысяче- → тысяча |
| 6 | - | "е" → "ь" | жизне- → жизнь |
| 7 | - | "о" → "ый" | крупно- → крупный |
| 8 | - | "о" → "ой" | криво- → кривой |
| 9 | - | "е" → "ий" | обще- → общий |
| 10 | "к/г" | "о" → "ий" | высоко- → высокий |
| **Inflexion rules** | | | |
| 11 | - | "ый" → "" | -этажный → этажн → этаж |
| 12 | - | "ий" → "" | -дилерский → дилерск → дилер |
| 13 | - | "ой" → "" | -звуковой → звуков → звук |

# Results for German language

| | Dictionary | Dictionary + Rules + Similarity | Dictionary + Rules + Similarity + Corpus | Banana Split | IMS Splitter [Weller and Heid, 2012] |
|---|---|---|---|---|---|
| Top 1 | 66% | **91%** | **91%** | 86% | 87% |
| Top 5 | 66% | 94% | **95%** | - | 92% |

# Usage of specialized corpus

## Helpful

- "Unknown" words
  Repowering-Leitfaden = $\text{repowering}_{corpus}$ + $\text{leitfaden}_{dico,corpus}$ [guide]
- Improves the ranking
  Traktionsbatterie [traction battery]
  without corpus : trakt + ion + batterie 1.0; traktion + batterie 1.0
  with corpus : traktion + batterie 1.50; trakt + ion + batterie 1.25

## Helpless

- Affects the ranking
  Aussichtsplattform [observation deck]
  without corpus (Top1) : aussicht + plattform
  with corpus (Top1): aus + sicht + plattform

? Solution: corpus frequency → domain specificity [Ahmad et al., 1992]

# Results for Russian language

| | Dico | Levenshtein | | | | Prefix | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Dico + Rules + Similarity | | Dico + Rules + Similarity + Corpus | | Dico + Rules + Similarity | | Dico + Rules + Similarity + Corpus | |
| | | Small rules | Large rules | Small rules | Large rules | Small rules | Large rules | Small rules | Large rules |
| Top 1 | 35% | 62% | 75% | 76% | **84%** | 58% | 68% | 72% | 78% |
| Top 5 | 35% | 71% | 81% | 86% | **92%** | 69% | 80% | 90% | 92% |

# Corpus + Similarity vs Large rules-set

In some cases, the corpus compensates for the lack of rules

- электромагнитный [electromagnetic] =
  $\text{электро}_{NC}$ + $\text{магнитный}_{corpus}$
- rule 11 (магнитный) = магнитн
  $\text{similarity}_{Lev}$ (магнитн, магнит) = 0.86
  result : электромагнитный = электро + магнит

# Conclusions and perspectives

## Conclusions

- Positive impact of LSP corpus for specialized vocabulary
- Splitting module can be useful to deal with Russian (terminology)

## Perspectives

- Test the algorithm for other fields and languages
- To split or not to split?
- Evaluate the impact of the splitting on MT

Спасибо за внимание!

Thank you for your attention!