Identifying similar text documents

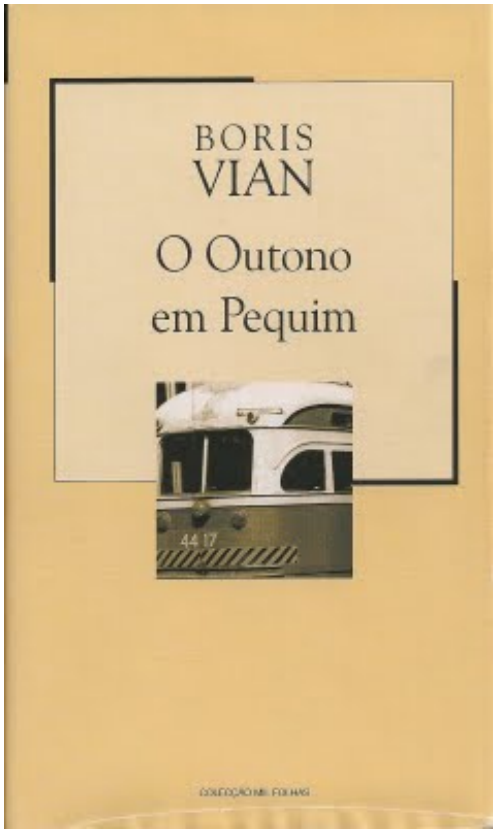
André Santos
andrefs@cpan.org

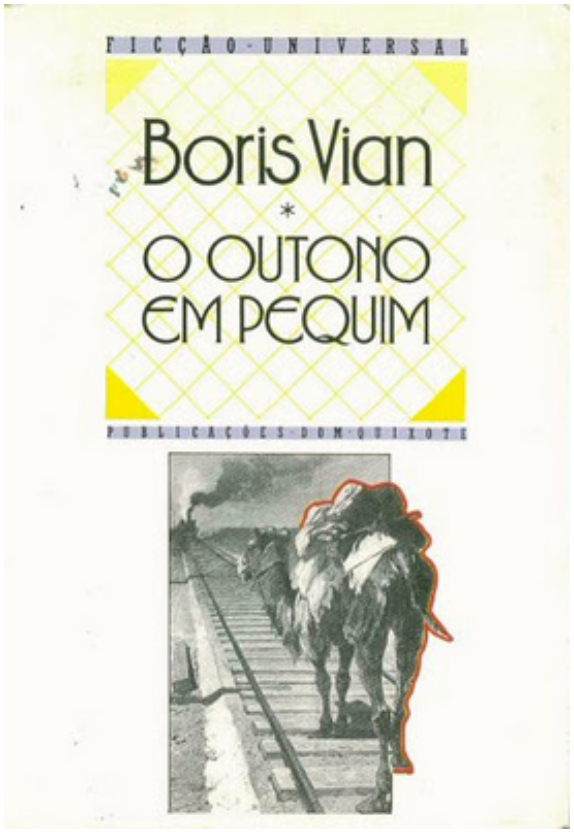
November 2011

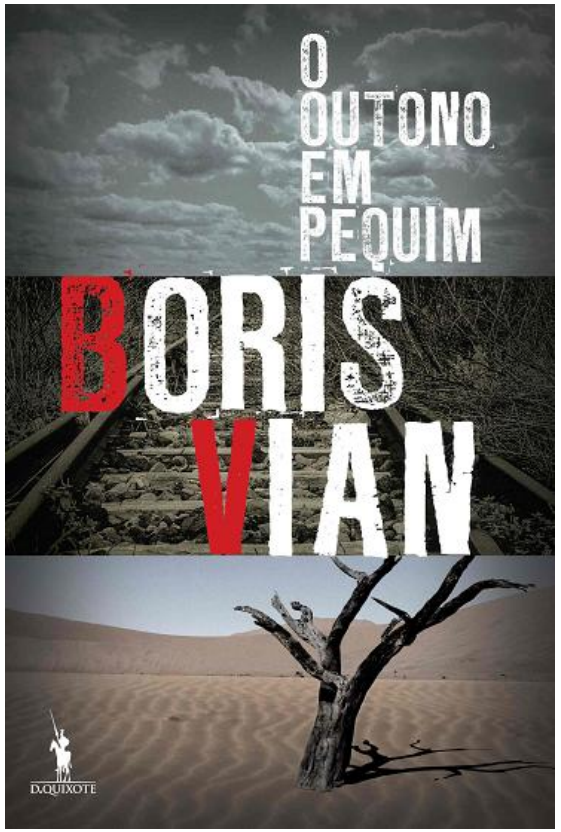
What we get


Mariah Stewart
The President's Daughter
LIKE A KNIFE
Temptation
The Bridal Veil
Lorraine Heath
The Outlaw Takes a Lady
JANE FEATHER
WEDDING GAME
Wedding Ring
Jennifer Crusie
DON'T LOOK NOW
Stephanie Laurens
A Rake's Vow
Sally MacKenzie
The Marquis

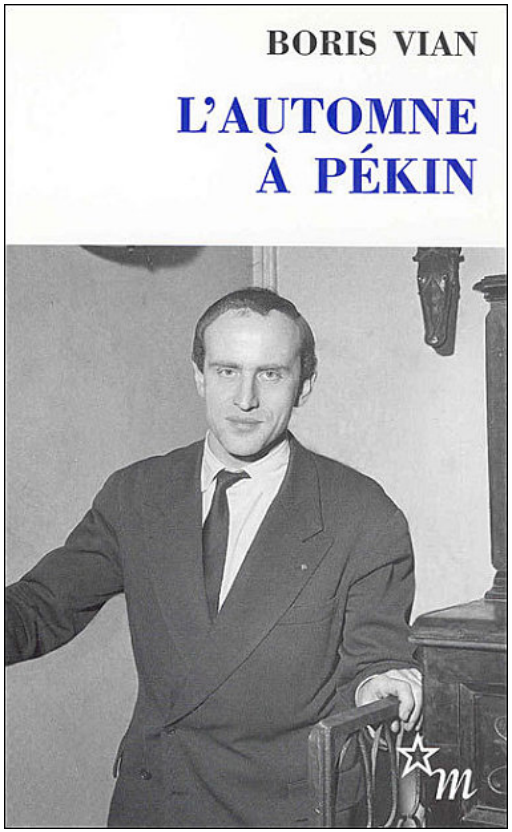
Duplicated versions

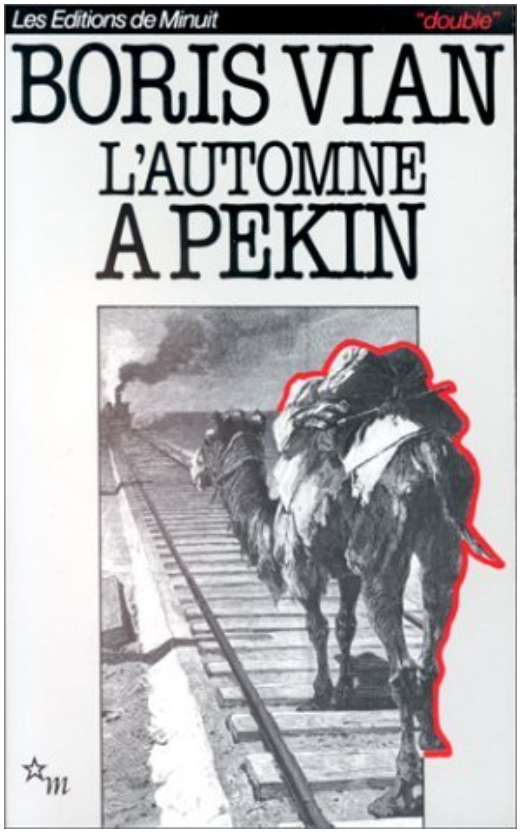




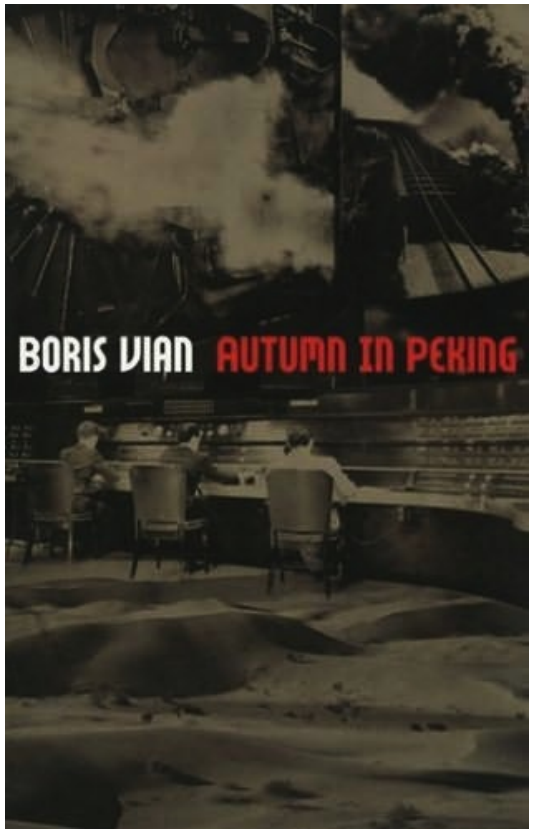


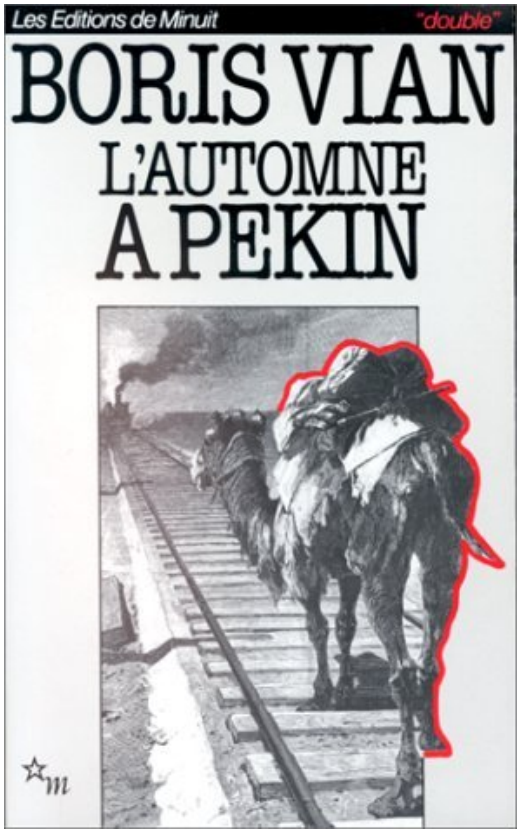
Duplicated versions

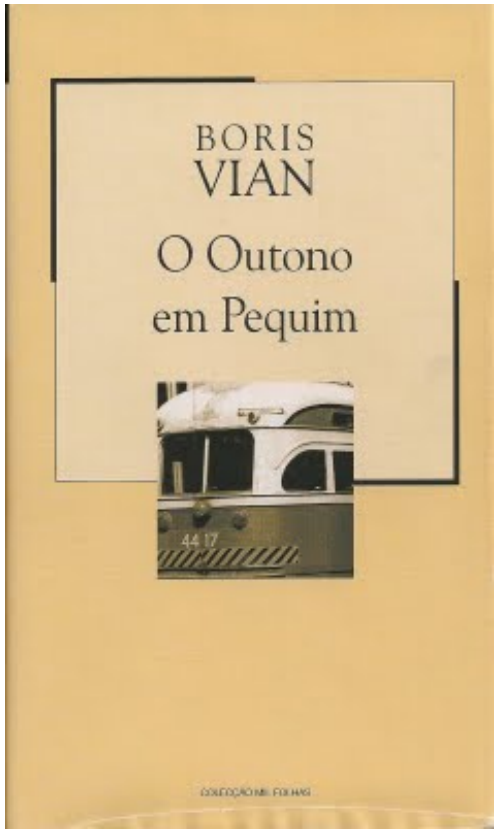




Candidate pairs

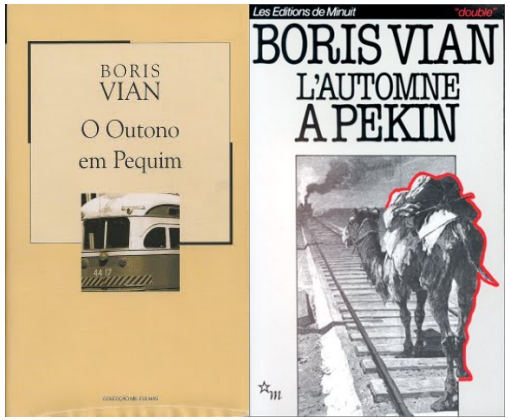


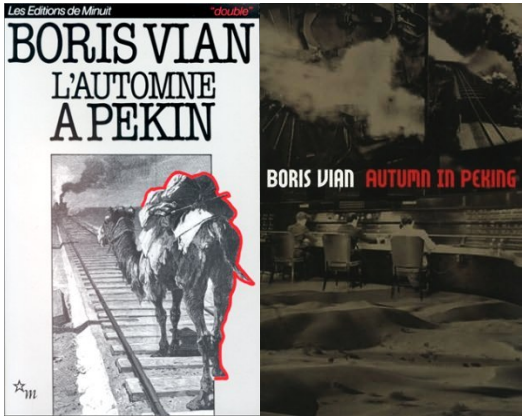




Candidate pairs

Candidate pairs





What this is really about

similarity

It's all LIEs!

Language Independent Element (LIE)

Terms which are usually kept untouched during translation.

It's all LIEs!

Language Independent Element (LIE)

Terms which are usually kept untouched during translation.

- Year references (e.g. "1977")

It's all LIEs!

Language Independent Element (LIE)

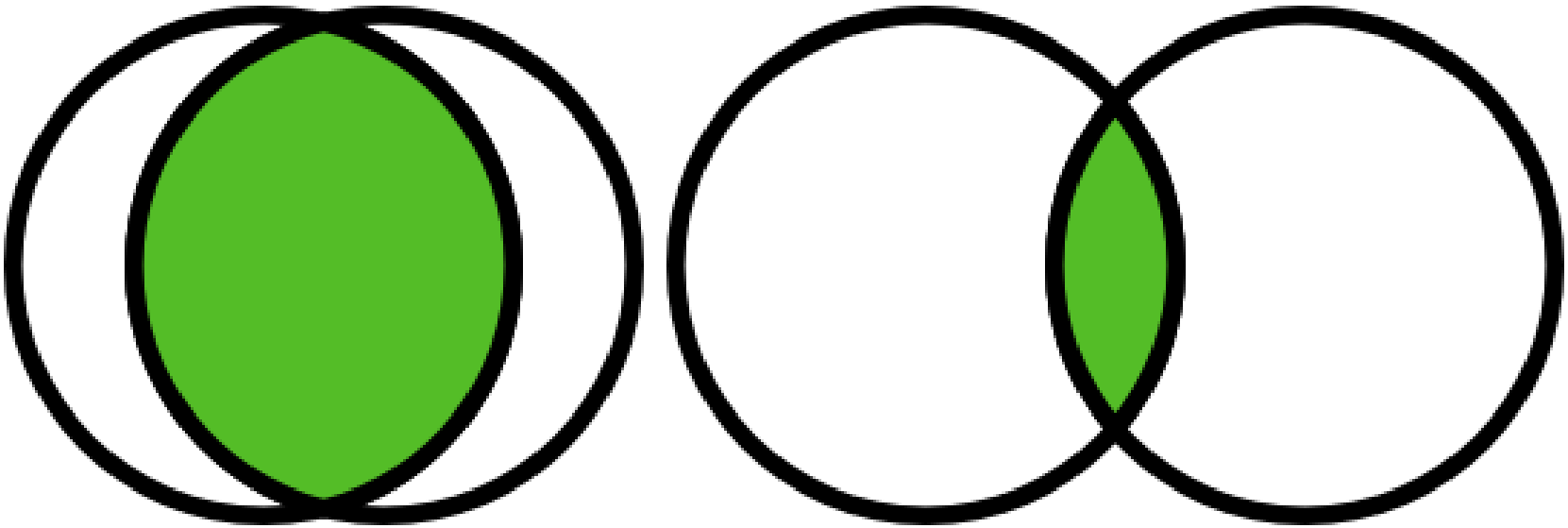
Terms which are usually kept untouched during translation.

- Year references (e.g. "1977")
- Proper names (e.g. "Sherlock Holmes")

Measuring similarity

$$similarity(A, B) = \frac{|A_{LIEs} \cap B_{LIEs}|}{|A_{LIEs} \cup B_{LIEs}|}$$

Measuring similarity



pairbooks

Similarity values

< 0.2 Documents are not related

> 0.4 Documents are candidate pairs

> 0.9 Documents are near duplicates

1.0 Documents are duplicates

Languages

High similarity, same language: (Near) duplicates

High similarity, different language: Candidate pairs

Behold, pairbooks!

```
~ $ pairbooks  PT_list.txt ES_list.txt

PTBR__Umberto_EcoO_nome_da_rosa.txt
  (0.227) [6954,7382]   ES__Umberto_EcoEl_Nombre_de_la_Rosa(...)
  (0.018) [6954,11408]  ES__Umberto_EcoEl_Pendulo_De_Foucau(...)
  (0.018) [6954,5604]   ES__Umberto_EcoDiario_Minimo__2.txt(...)

PTBR__Umberto_EcoO_Pendulo_de_Focault.txt
  (0.391) [11276,11408] ES__Umberto_EcoEl_Pendulo_De_Foucau(...)
  (0.042) [11276,6024]  ES__Umberto_EcoLa_busqueda_de_la_Le(...)
  (0.035) [11276,5604]  ES__Umberto_EcoDiario_Minimo__2.txt
(...)
```

Perfect LIEs do not exist

Year references

- Can be confused with page numbers
- Headers/footers can contain them (publishing year, copyright, . . .)

Proper names

- Sometimes are translated (e.g. “São Tomé”, “Judas Tomé”, etc)
- Some languages use different scripts (e.g. Russian)
- Some languages have declensions

. . .

How to improve LIEs (future work)

- accept a list of equivalent words
- accept a list of stop words
- ...

Give me one of those!

CPAN

`http://search.cpan.org/perldoc?pairbooks`

- Developer version
- requires Linux, Perl
- Incomplete documentation

Identifying similar text documents

André Santos
andrefs@cpan.org

November 2011