# Issues of Multilingual Topic Generation

Christa Womser-Hacker

Universität Hildesheim, Germany

CLEF 2001, 4.09.01

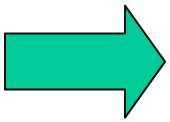
# Content

- Comparison with TREC
- New Challenges for CLIR
- „Rules“ for Topic Generation
- Process of Topic Generation
- Problems
- Summary

# Topic Generation à la TREC

- Topics are original "user" requests
- Topic builders judge their relevance
- Topics give criteria for relevance
- Topics have changed over the TREC programme
  - TREC 1 and 2: very carefully formulated
  - From TREC 3 onwards: shorter, less elaborated, more realistic for the adhoc task
  - Control of the effects of topic styles (long, short, very short)
- Reflections on Topic generation is necessary

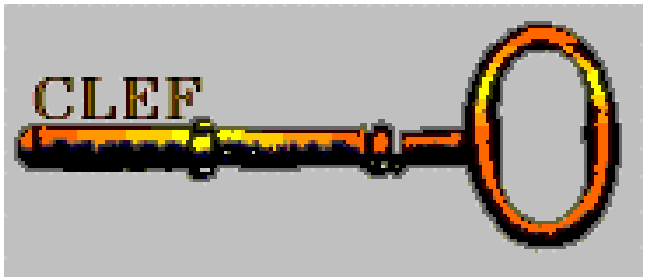
# Topic Generation in a Multilingual Environment



New issues, new problems

## Starting Questions

- How to get real user requests?
- Could real user requests simply be translated?
- Who judges the relevance?
  - cooperative work, a lot of discussions



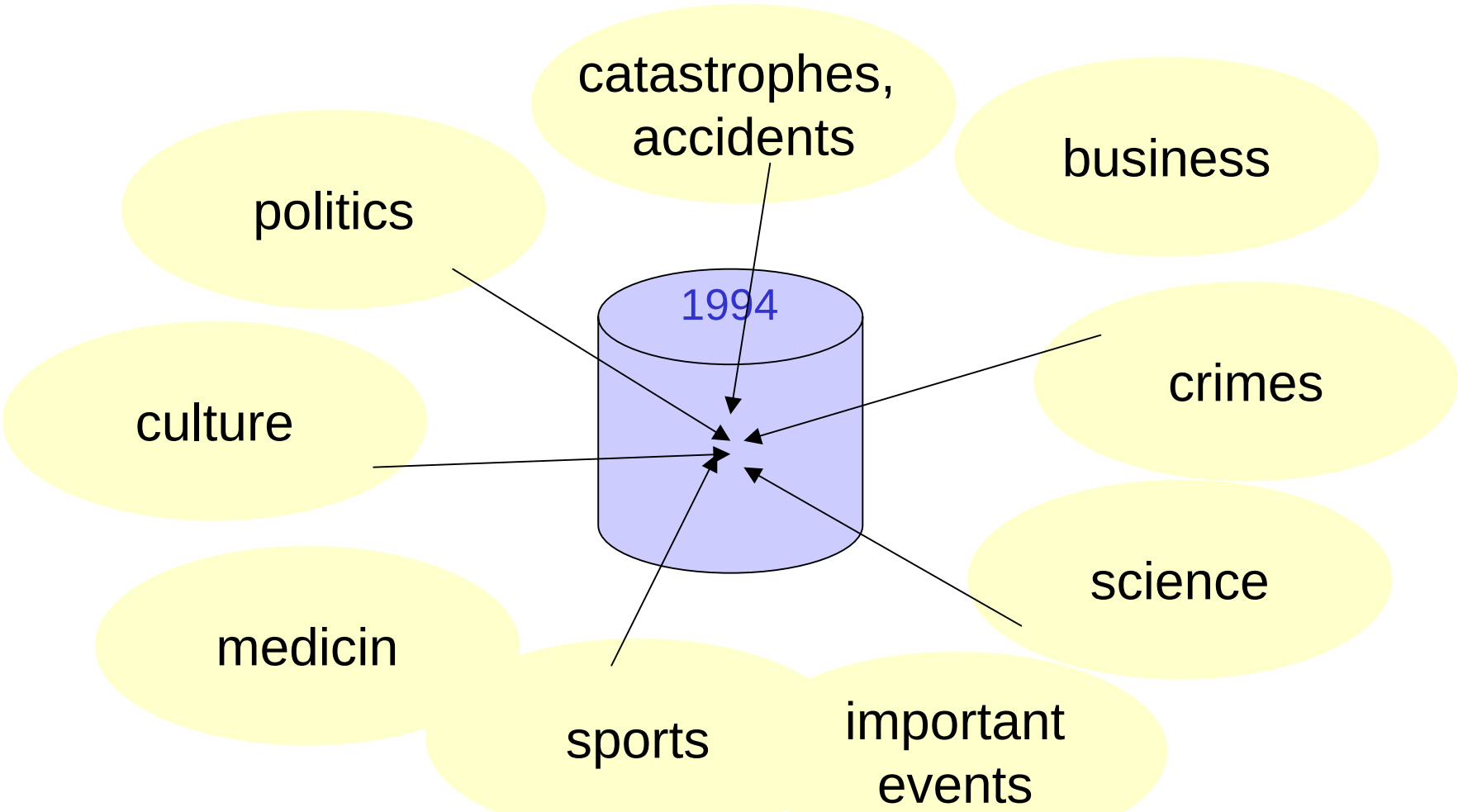

...the only restriction

# Documents

## 1994 newspapers

Los Angeles Times
Le Monde
La Stampa
NZZ / SDA
Frankfurter Rundschau / Der Spiegel
Spanish Newswire ?
etc.

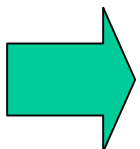
# What happened in 1994? What was interesting?




Resources: Almanachs, Chroniken, Year Books, Le Monde hébdomadaire

# Topic Generation Process

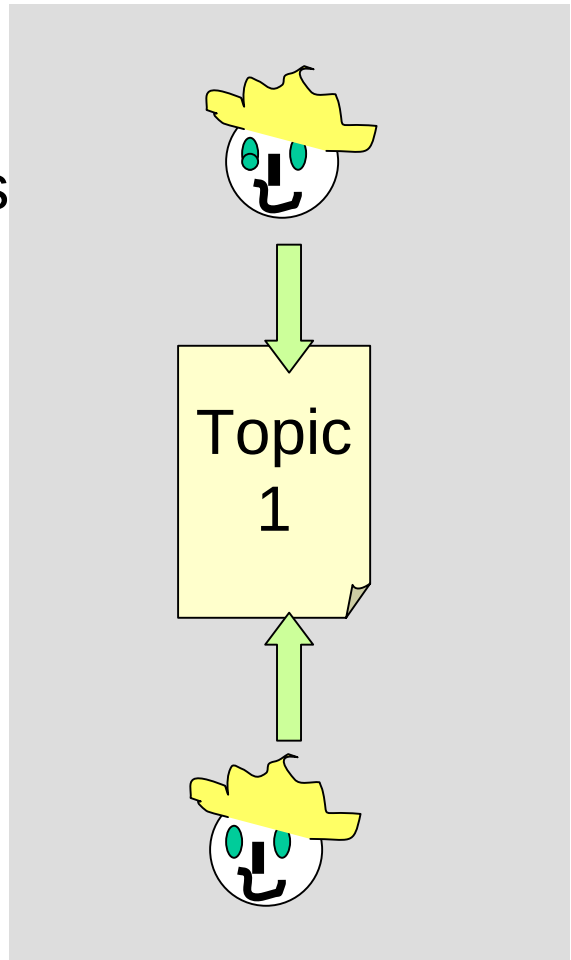
- Suggestions for topics in each language / culture
- Potential topics were retrieved in each language
- Discussion and selection
- Generation of the English master set
- Check the quality of the master set
- Translation of the topics
- Check of the translations
- Discussion of modifications with the topic builders



50 topics in each language

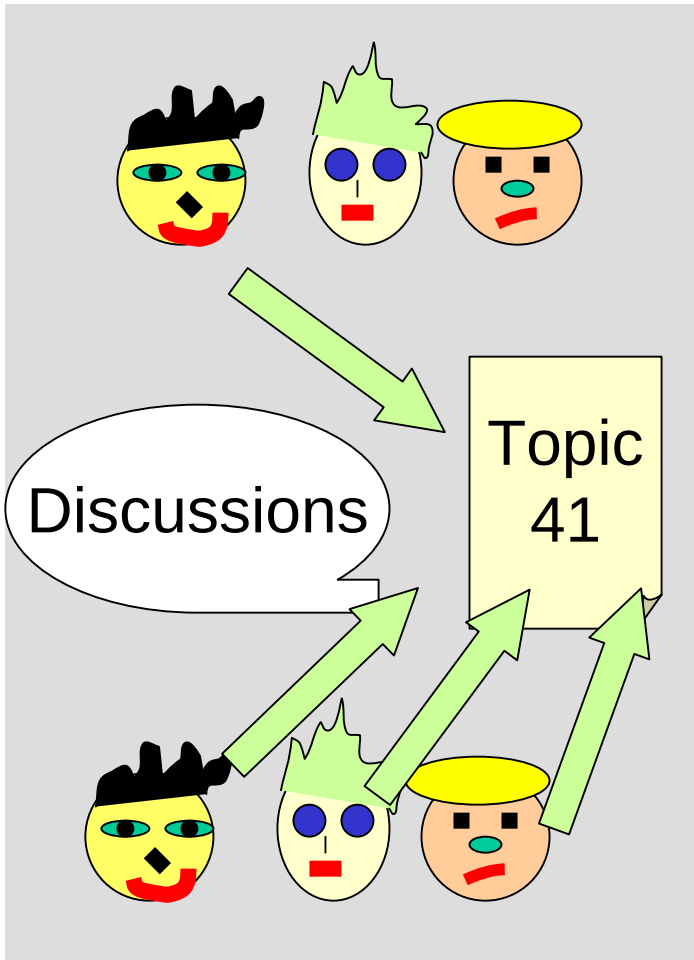
# TREC

Topic Generators



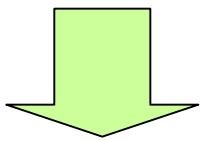

Relevance Assessors

# CLEF



# 50 Original Topics

- 9 English original topics
- 8 Italian, French, German, Spanish original topics
- 5 Dutch original topics
- 4 Japanes original topics



translated to English

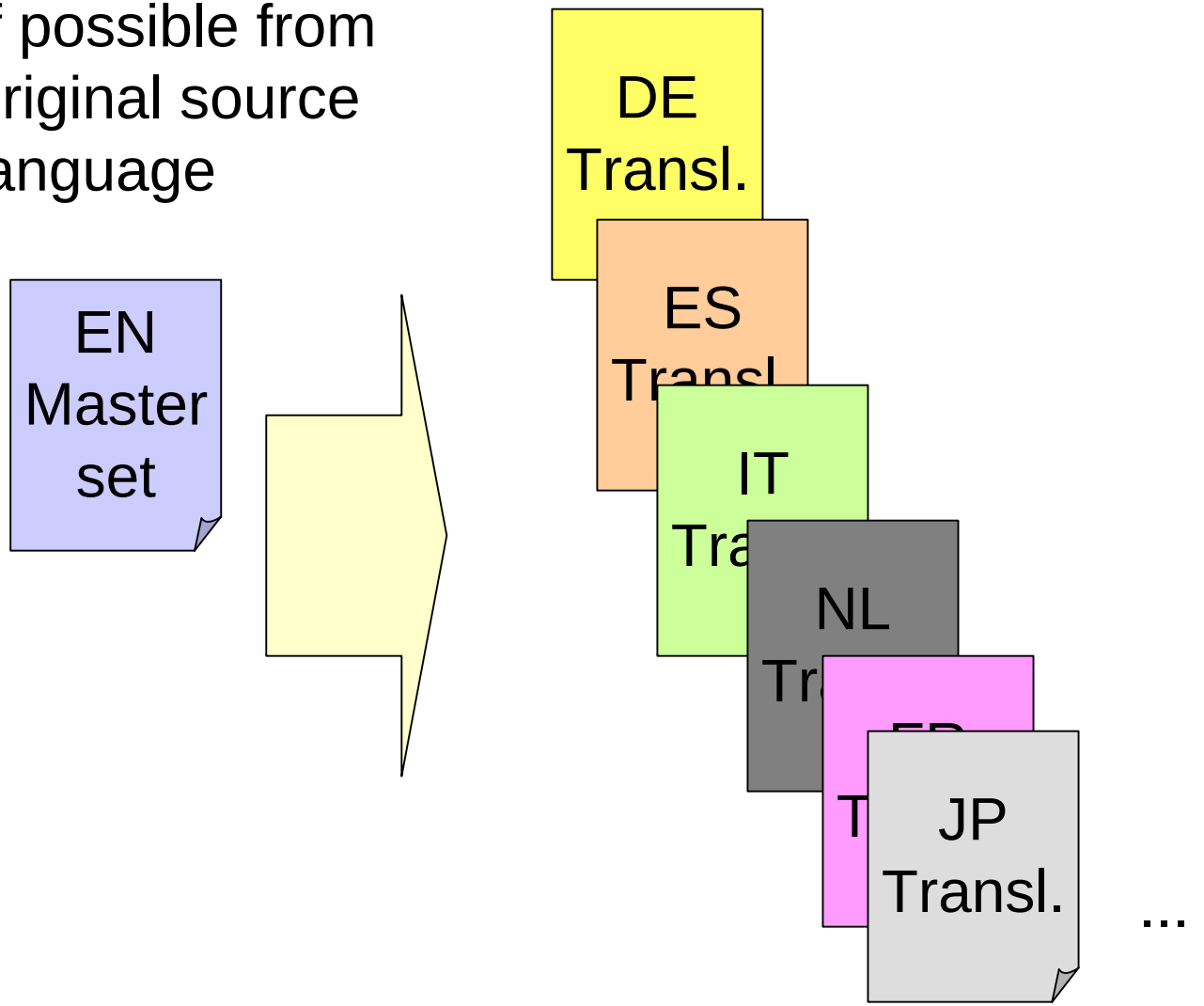
- English Master set

# Topic Generation Rules

- General rules (in each language) or better desiderata:
  - Contents of Topics:
    - 1/3 international, 1/3 european, 1/3 national
  - 20 % proper names, events, facts
    - Query-answering
  - There should be documents in all newspapers (no max. limit)

# Topic Translation

If possible from original source language



# Example

<top lang="EN">
<num>1</num>
<EN-title>Shark Attacks</EN-title>
<EN-desc>Documents will report any information relating to shark attacks on humans.</EN-desc>
<EN-narr>Identify instances where a human was attacked by a shark, including where the attack took place and the circumstances surrounding the attack. Only documents concerning specific attacks are relevant; unconfirmed shark attacks or suspected bites are not relevant.</EN-narr></top>

# Translation of topics in all languages

***Goal***: All systems should have same conditions

- Transfer the texts in another linguistical and cultural context
- Term semantics not isolated but from the term position within a system

  Ambiguities!
- Not always 1:1 translation possible

Systems should meet these challenges

# Problems found in the topics

- Compound words
- Proper names
- Abreviations
- Phrases
- Idomatics / metaphors
- Domain specific terminology
- Culture specific knowledge involved
- Query-answering topics and open requests
- Remarks on relevance in NARR

# Further Problems

- American – British English
  - (real) estate manager
- Swiss and German German
  - Müllverbrennung – Kehricht...
- German "Rechtschreibreform„
  - Rad fahren, radfahren
  - Prozessakten, Prozeßakten

# Topic Check

- Correctness of translations
  - Only by native speakers and (technical) translators
  - If language pair not available, deviation by English master set
- Check of formal tags
- Problem discussion
- Analysis of modifications

# Topic Check

- Synoptical view of the 50 topics
- Check of translations
  - Different people and tools
- Look for contradictions (same topic, exactly the same facts)
- Discussion with topic generators
- What will be relevant? (see narr.)
- Random ordering of topics

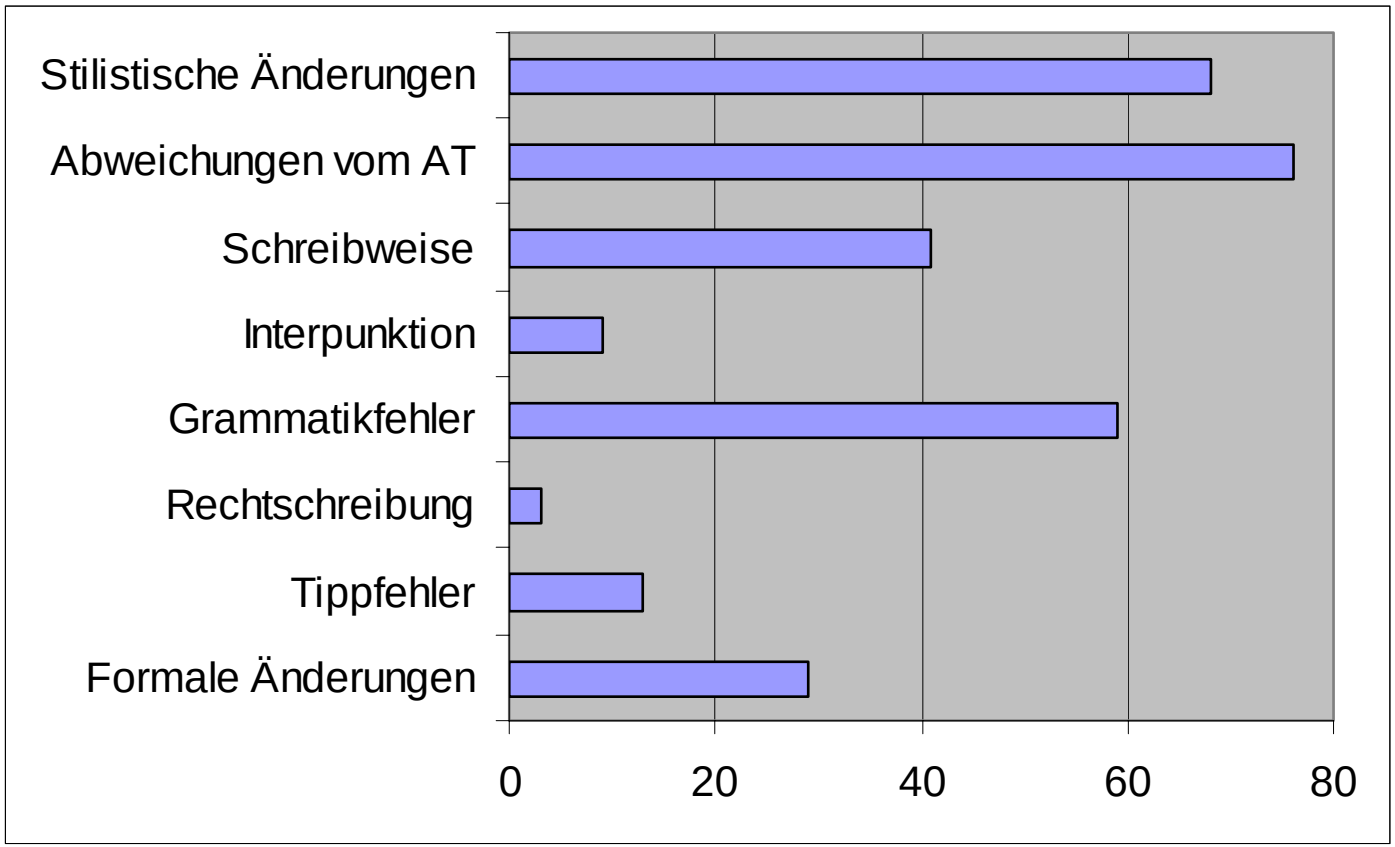
# Categorization of modifications (1)

- Formal changes
- Typos
- Orthography
- Interpunction
- Grammar
  - Preposition
  - Tempus
  - Numerus
  - Modus
  - Gender

# Categorization of modifications (2)

- Language specials
  - Compounds
  - US / UK English
  - German / Swiss German
  - Proper names
  - Upper / lower case
  - Semantical changes
  - Ellipses
  - Add-ons
  - Style

# Categorization of modifications




Quelle: PK Hildesheim SoSe 2001

# Example: political correctness

<top lang="DE">
<num>C050</num>
<DE-title> Aufstand in Chiapas </DE-title>
<DE-desc> Berichte über die Revolte von **Indios** in Chiapas (Mexiko) sind gesucht. </DE-desc>
<DE-narr> Die Dokumente berichten über die Hintergründe und den Verlauf des Aufstands der **indigenen Bevölkerung** im mexikanischen Chiapas. Sie stellen auch die Reaktionen der mexikanischen Regierung dar. </DE-narr>
</top>

# Example: „Muisarm“

<num>C064</num>
<NL-title> Muisarm </NL-title>
<NL-desc> Zoek documenten waarin melding wordt gemaakt van een muisarm.
</NL-desc>

<num>C064</num>
<FR-title>Ordinateur: souris et tensions musculaires</FR-title>
<FR-desc>Rechercher des documents sur les lésions dues aux mouvements répétitifs (RSI repetitive strain injuries) provoquées par l´utilisation de la souris de l´ordinateur.</FR-desc>

# Example: „Renewable Power“

<top lang=“EN“>
<num>C086</num>
<EN-title>Renewable Power</EN-title>
…
<EN-narr>Relevant documents discuss the use of renewable energy sources such as solar, wind, biomass, hydro, and geothermal sources. Low emission vehicles as for example electric or CNG cars are not relevant. Fuel cells are not relevant unless their fuel qualifies as renewable.</EN-narr>
</top>

<top lang="EN">
num>C086</num>
<DE-title> Erneuerbare Energien </DE-title>
<DE-narr> Relevante Dokumente behandeln die Nutzung erneuerbarer Energiequellen, wie der Sonne, des Windes, der Biomasse, des Wassers und der Erdwärme. Schadstoffarme Fahrzeuge wie z.B. Elektroautos oder mit Flüssiggas betriebene Autos sind irrelevant. Brennstoffzellen sind nicht relevant, solange der Brennstoff nicht als erneuerbar gilt. </DE-narr>
</top>

# Example: „Lennon“

<top lang="ES">
<num>C083</num>
<ES-title> Subasta de objetos de Lennon. </ES-title>
<ES-desc> Encontrar subastas públicas de objetos de John Lennon.</ES-desc>
<ES-narr> Los documentos relevantes hablan de subastas que incluyen objetos que pertenecieron a John Lennon, o que se atribuyen a John Lennon.</ES-narr>
</top>

<top>
<num>C083</num>
<FR-title> Vente aux enchères de souvenirs de John Lennon </FR-title>
<FR-desc> Trouvez les ventes aux enchères publiques des souvenirs de John Lennon. </FR-desc>
<FR-narr> Des documents pertinents décriront les ventes aux enchères qui incluent les objets qui ont appartenu à John Lennon ou qui ont été attribués à John Lennon. </FR-narr>
</top>

# Example: „Schneider“

<top lang="DE">
<num>C089</num>
<DE-title> Schneider-Konkurs </DE-title>
<DE-desc> Konkurs des deutschen Immobilienhändlers Schneider. </DE-desc>
<DE-narr> Die Dokumente berichten über den Konkurs des deutschen Immobilienhändlers Schneider und dessen Hintergründe. Sie untersuchen auch die Unterlassungen, Fehler und Verantwortlichkeit der deutschen Banken in diesem Fall. </DE-narr>
</top>

<top>
<num> C089</num>
<FR-title> Faillite de M. Schneider</FR-title>
<FR-desc> Faillite de l’agent immobilier allemand Schneider</FR-desc>
<FR-narr>Les documents pertinents donnent des informations sur la faillite de l’agent immobilier allemand Schneider et sur les raisons de cette faillite. Ils prennent aussi en considération les omissions, les erreurs et la responsabilité des banques allemandes dans cette affaire. </FR-narr>
</top>

# Summary

- Topic generation is important but hard work
- Further analyses are necessary
  - What makes a topic difficult?
  - Which facilities should systems have?
  - Correlation: topic – r&p values
  - ...


EN DE FR EN NL JP FR EN IT FR
FR ES FR DE JP FR ES IT EN IT
IT NL DE ES IT FR EN IT EN
ES EN NL EN DE NL IT ES NL DE
JP ES ES EN FR ES JP DE DE IT
Thanks for your attention!