Pluricentric languages and Wikipedia – where to draw the line... and where not to

Philip Kopetzky (philip.kopetzky@gmail.com)
User:Braveheart, member of Wikimedia Austria
July 19th, Wikimania 2015
Introduction

• Term “pluricentric language” coined by William Stewart and Heinz Kloss in the late 1960s

• Definition of a pluricentric language (Clyne, 1992)
  • Several standard versions
  • Used across boundaries of individual political entities

• Almost all big languages are pluricentric
  • German, English, French, Spanish, Chinese, Hindustani, etc.

• With some notable exceptions
  • Russian, Japanese, Albanian

• Differentiation between asymmetrical and symmetrical pluricentric languages
Asymmetrical pluricentric languages

- Found wherever there is a dominant variety of the language.
- Mostly due to a very large percentage of native speakers located in one country.

“The norms of one national variety (or some national varieties) is afforded higher status than other varieties, internally and externally, than those of the others” (Clyne, 1992)

- In some cases cultural influence of the dominant variety threatens the existence of smaller varieties.
Examples in detail: German

“What divides Germans and Austrians is their common language” - unknown

• Variants:
  • German Standard German (D)
  • Austrian Standard German (A)
  • Swiss Standard German (CH)

• Distribution
  • D: 77 million (86 %)
  • A: 7.5 million (8 %)
  • CH: 5 million (6 %)
Asymmetrical examples in detail: German (2)

• Highly standardised language
  • Rat für deutsche Rechtschreibung (Council for German Orthography)
    • Includes councillors from Germany, Austria, Switzerland, South Tyrol, Belgium and Liechtenstein
  • Duden as leading authority in Germany
  • Official “Austrian dictionary” for Austria and South Tyrol (since 1951)
  • “Swiss society for the german language” in Switzerland

• Several straw polls in Wikipedia tried to harmonise variants, but failed

• Typography and grammar in articles based on ties to a country ([[de:Wikipedia:Österreichbezogen]], [[de:Wikipedia:Schweizbezogen]])
“Januar” vs. “Jänner”

• German for January

• “Jänner” as Austrian variant vs. “Januar” in Germany and Switzerland

• Straw poll in 2005 decided to keep the Austrian variant, limiting it to articles with ties to Austria

• Constant changes by German users from “Jänner” to “January”

• Despite having official guidelines ([:de:Wikipedia:Datumskonventionen]], number of changes stay the same over the years
Number of changes which state "Jänner" as reason for an edit by month

(Source: dewiki_p on Labs DB)
Why do conflicts arise to this day?

• Documentation of the guidelines not sufficient?
  • Traffic for those guidelines at around 8 hits per day in June 2015

• Knowledge of German as a pluricentric language not widely spread?
  • Objection raised that everything besides German Standard German is a dialect comparable to inner German regional dialects and therefore not compatible with Standard German
Examples in detail: French

- Francophone world
Examples in detail: French (2)

• About 80 million native speakers
• Dominant variant is the standardised variant of French in France
  • Controlled by the “Académie française“ (French Academy), founded in 1635
• Other standardised variants
  • Belgian French
  • Canadian French (regulated by the Office québécois de la langue française)
• Use of design theory „Principle of least astonishment“ (POLA) in French Wikipedia also applies to wording in articles
A tale of chicory and ice hockey pucks

- Chicory: Commonly called “endive” in French, “chicon” used in Northern France and Belgium
- Settled by a straw poll in February 2007 regulating article names in the field of biology
- Arguments: POLA and a table comparing the use in various francophone countries combined with their population, among other arguments
A tale of chicory and ice hockey pucks (2)

• Ice hockey pucks: Called “palet” in francophone Europe, “rondelle” in Quebec

• Modern ice hockey was invented in Quebec, therefore it has strong ties to this region

• POLA and the overwhelming majority that use “palet” decide this issue, accompanied by this table:

<table>
<thead>
<tr>
<th>Pays (Country)</th>
<th>Population francophone</th>
<th>Palet</th>
<th>Rondelle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Québec</td>
<td>7 600 000</td>
<td>OUI</td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>64 000 000</td>
<td>OUI</td>
<td></td>
</tr>
<tr>
<td>Belgique</td>
<td>4 200 000</td>
<td>OUI</td>
<td></td>
</tr>
<tr>
<td>Suisse</td>
<td>1 500 000</td>
<td>OUI (or “puck”)</td>
<td></td>
</tr>
<tr>
<td>Totaux</td>
<td>77 300 000</td>
<td>69 700 000</td>
<td>7 600 000</td>
</tr>
</tbody>
</table>
What does this comparison tell us?

• Arguments for and against the use national variants in asymmetrical languages in Wikipedia
  • Standardising the language in Wikipedia according to the majority of readers makes it easier for most readers
    • ... but does this apply to all articles? What about local themes? (Villages, regional politicians, etc.)

• Fostering diversity in Wikipedia makes for happier users
  • ... and even dictionaries in Germany are starting to accommodate other standard variants

• Demotivating users from small user groups sounds like a bad idea
  • Who else is going to write about those regions in detail?
Symmetrical pluricentric languages

• Found wherever there are two or more centres who each use their own widely spread, standardised variant

• Often a product of the “heartland” of the language as one centre and another country with a large population as the other centre (Ammon, 2004)
  • Examples: Portuguese, English, Spanish
Examples in detail: English

- WP:TIES

Rebozo patterned ties,
FeriadeRebozo2014 48.JPG,
AlejandroLinaresGarcia,
CC-BY-SA 4.0
Examples in detail: English

• WP:TIES
Examples in detail: English (2)

• “England and America are two countries divided by a common language.” – Georg Bernard Shaw

• WP:TIES
  • Defined as “An article on a topic that has strong ties to a particular English-speaking nation should use the English of that nation”.
  • At the same time, no claim to any article according to [[en:Wikipedia:Ownership of articles]]
  • Grey / uncertain areas are common whenever two or more countries have strong ties to a topic
    • For example a Sherlock Holmes series produced in the United States
Examples in detail: English (3)

• Manual of style
  • Opportunities for commonality
  • Consistency within articles
  • Strong national ties to a topic
  • Retaining the existing variety

• Sheer number of rules makes for a confusing interpretation of these guidelines
  • Most terms with differing vocabulary (like [[en:center]] and [[en:elevator]]) are written in American English
Spanish

• Use of the term “Castellano” in South America instead of “Idioma español”

• Political motives as reason for using one over the other?

• Traffic ratio of redirect “Castellano” to article “Idioma español” approx. 1:50 on Spanish Wikipedia (June 2015, Source: stats.grok.se)
Outlook

- Analysis of Wikipedia versions of regional dialects that don’t have a standardised written form

- Statistical analysis of article traffic based on the country/region of readers

- Viability of pluricentric languages with more than one Wikipedia version (e.g. Serbo-Croatian)
References / Further reading


• Wikipedia ([de,en,es,fr](https://en.wikipedia.org/wiki/Pluricentricity))
Discussion

Questions?
Comments?
Own experiences on this matter?