Translation of *you* in online film subtitles:

A quantitative cross-linguistic study

Natalia Levshina
University of Leipzig
Object of study

• T/V-distinction in addressing the hearer
• The distinction is present in most European languages
  • T forms: informal, familiar, e.g. French *tu*, German *du*,
    Russian *ty* + Verb 2\textsuperscript{nd} SG
  • V forms: formal, polite, e.g. French *vous*, German *Sie*,
    Russian *vy* + Verb 2\textsuperscript{nd} PL or 3\textsuperscript{rd} SG/PL
Cross-linguistic research

- WALC Chapter 45, Helmbrecht 2013

Values

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>No politeness distinction</td>
<td>136</td>
</tr>
<tr>
<td>Binary politeness distinction</td>
<td>49</td>
</tr>
<tr>
<td>Multiple politeness distinctions</td>
<td>15</td>
</tr>
<tr>
<td>Pronouns avoided for politeness</td>
<td>7</td>
</tr>
</tbody>
</table>
Research questions

• What are the cross-linguistic (dis)similarities wrt. the relative frequencies of the forms?
• What are the cross-linguistic (dis)similarities wrt. the preferences of the forms in different communicative situations?
Power and solidarity (Brown and Gilman 1960)

• Power dimension:
  • Based on “older than”, “richer than”, “parent of”, etc.
  • Systematic distinction from the late Middle Ages. Everyone has his/her fixed place in the society.

• Solidarity dimension:
  • Based on “the same age/family/class as”.
  • Emerged with social mobility and egalitarian ideology. Starting from the French revolution (*Citoyen, tu*).
  • Currently dominates in major European languages, but there are subtle cross-linguistic differences.
Languages in the sample

• Germanic: Dutch, German and Swedish
• Romance: French and Spanish
• Slavic: Bulgarian, Polish and Russian
• Greek
• Finnish
# T/V forms (standard varieties)

<table>
<thead>
<tr>
<th></th>
<th>Nr of types</th>
<th>T-pronoun</th>
<th>V-pronoun(s), one person</th>
<th>V-verb agreement, one person</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>2</td>
<td>du</td>
<td>Sie</td>
<td>3rd person PL</td>
</tr>
<tr>
<td>Dutch</td>
<td>2</td>
<td>jij (je)</td>
<td>u</td>
<td>2nd person SG</td>
</tr>
<tr>
<td>Swedish</td>
<td>2</td>
<td>du</td>
<td>ni</td>
<td>2nd PL</td>
</tr>
<tr>
<td>French</td>
<td>2</td>
<td>tu</td>
<td>vous</td>
<td>2nd PL</td>
</tr>
<tr>
<td>Spanish</td>
<td>2</td>
<td>tú</td>
<td>usted</td>
<td>3rd person SG</td>
</tr>
<tr>
<td>Russian</td>
<td>2</td>
<td>ты [ty]</td>
<td>вы [vy]</td>
<td>2nd PL</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>2</td>
<td>ти [ti]</td>
<td>Вие ['vi.ɛ]</td>
<td>2nd PL</td>
</tr>
<tr>
<td>Polish</td>
<td>2</td>
<td>ty</td>
<td>pan (m)/pani (f)</td>
<td>3rd person SG</td>
</tr>
<tr>
<td>Greek</td>
<td>2</td>
<td>εσύ [e'si]</td>
<td>εσείς [e'sis]</td>
<td>2nd PL</td>
</tr>
<tr>
<td>Finnish</td>
<td>2</td>
<td>sinä</td>
<td>te</td>
<td>2nd PL</td>
</tr>
</tbody>
</table>
Data: ParTy corpus

• A Parallel corpus for Typologists
• Online subtitles of films and TED talks in many languages simultaneously
• Mostly Indo-European, but also Chinese, Turkish, Finnish, Indonesian, Vietnamese...
• Partly available from www.natalialevshina.com/corpus.html
Why subtitles?

Cluster Dendrogram

Based on the frequencies of 3-grams (Levshina, Forthcoming)
Films
Data set

- English data: instances of *you/yourself* used when referring to one person.
- 243 communicative situations with unique participants (in order to ensure maximal diversity)
- Translations into 10 languages coded for T or V
Variables describing relationships between Speaker and Hearer

- **Rel_Age**: is H younger, older or of the same age (approximately) than/as S?
- **Rel_Class**: does H belong to a higher, lower or same social class as S?
- **Rel_Power**: does H have social power over S? E.g. employer > employee, prime-minister > minister, general > soldier
- **Rel_Gender**: M to M, M to F, F to M, F to F
- **Rel_Circle**: family, romance, friends, working together, school or university, “house” (household servants, hotel), prison, acquaintances, strangers
Variables describing Speaker and Hearer

• **S_Age**: age of S (child, young person, middle-aged, elderly)
• **H_Age**: age of H (child, young person, middle-aged, elderly)
• **S_Class**: social class of S (upper, middle, lower, other)
• **H_Class**: social class of H (upper, middle, lower, other)
• **S_Gender**: gender of S (M or F)
• **H_Gender**: gender of H (M or F)
Variables describing communicative settings

- **Others**: are there other people involved in the situation?
- **Office**: does the interaction take place in an office, a government building, prison, school, etc.?
- **Before68**: does the action take place before 1968?
- **Britain**: does the action take place in Great Britain?
Example from *The Grand Budapest Hotel*

*M. Gustave*: What have you done to your fingernails?

*Madame D*: I beg your pardon?

*M. Gustave*: This diabolical varnish. The color is completely wrong.

*Madame D*: Don’t you like it?

*M. Gustave*: It’s not that I don’t like it. I am physically repulsed.
Coding of the situation

• Relational variables:
  • \textit{Rel\_Age}: H is younger
  • \textit{Rel\_Class}: H belongs to a lower social class
  • \textit{Rel\_Power}: H has less power
  • \textit{Rel\_Gender}: F to M
  • \textit{Rel\_Circle}: romance

• Individual variables:
  • \textit{S\_Age}: elderly
  • \textit{H\_Age}: middle
  • \textit{S\_Class}: upper
  • \textit{H\_Class}: lower
  • \textit{S\_Gender}: F
  • \textit{H\_Gender}: M

• Settings:
  • \textit{Others}: no
  • \textit{Office}: no
  • \textit{Before\_68}: yes
  • \textit{Britain}: no
Example of translations

• EN: Mal, what are you doing here? (Inception)
  • DE: Mal, was tust du hier? (T-pronoun and T-verb form)
  • RU: Мол, что ты здесь делаешь? (T-pronoun and T-verb form)
  • ES: Mal, qué haces aquí? (T-verb form)
  • BG: Какво правиш тук? (T-verb form)
Proportions of T/V in 10 languages
Conditional inference trees

• deal with any data, continuous or categorical; can be an alternative to most popular types of generalized linear models

• based on binary recursive partitioning:
  1. The algorithm is in search of the covariate X that is the most strongly associated with the response Y (e.g. has the smallest $p$-value or test statistic).
  2. The algorithm decides on the best way of splitting the data into two subsets with different values of X.
     e.g. if X has values a, b and c, one can split it into
     [a, b] and c
     a and [b, c]
     [a, c] and b
  3. For each subset, repeat 1 and 2 until certain conditions are met (e.g. there is at least one $p < 0.05$). If not, stop.
Conditional inference tree: Swedish
Conditional inference tree: German
Random forests

• Aggregate data from many trees (e.g. 500 or 1,000).
• Trees are created from bootstrapped data (a smaller sample is drawn randomly from the original data).
• Improved classification accuracy in comparison with individual trees (adjusts the instability of individual trees)
• Conditional variable importance: how important each predictor is, given the other predictors and possible interactions with them.
  • For each tree: compare the classification accuracy with the given predictor X before and after permutation.
  • The greater the decrease in the classification accuracy, the more important X.
  • Conditional: while controlling for the values of the other predictors (see Strobl et al. 2008)
  • Average this information across the trees.
Random forest: German

Conditional variable importance

Rel_Circle  H_Age  Office  Sex_H  Rel_Gender  S_Age  Britain  Rel_Class  Before68  H_Class  Rel_Power  S_Class  Rel_Age  Others  Sex_S
Random forests: Germanic

Conditional variable importance
Random forests: Romance

Conditional variable importance
Random forests: Slavic

Conditional variable importance

- Bulgarian
- Polish
- Russian
Random forests: Greek and Finnish

Conditional variable importance
Results

• The most important variable is *Rel_Circle* (no. 1 in most languages, no. 2 in some):
  • In all countries, strangers and acquaintances (low intimacy) are more frequently addressed by V, and family and friends (high intimacy) as T. There is variation, as far as other levels of intimacy are concerned. E.g. in German and French, the T forms are restricted to friends and family. In Finnish, Greek and Polish, they are the least restricted.

• The second most important variable is relative social class:
  • Usually, when H is higher on the social ladder than S, he/she is addressed by V.
  • Conditional on time (in Swedish, before 1968) and circle (in Bulgarian and Finnish, only in higher-intimacy contexts).
Results

- **Rel_Age**: older H are more frequently addressed by V in low-intimacy contexts in Polish
- **Rel_Gender**: men speaking to men in low-intimacy situations prefer V in Finnish
- **S_Age**: younger S use T in low-intimacy situations in Bulgarian
- **H_Age**: young Hearers are addressed more frequently by T in lower-intimacy situations in French
- **Office**: more V in the office in German
- **Before68**: more V in French (for younger H in low-intimacy situations) and Swedish
- **Britain**: more V between equals or youngsters in low-intimacy contexts in Bulgarian, Greek, Polish and Spanish
Conclusions

• The solidarity dimension is indeed the most universal and powerful, although there is substantial variation regarding the boundary between high and low intimacy.

• The power dimension is still present, however.

• There are more constraints and cross-linguistic variation in low-intimacy contexts. Higher ‘embarrassment potential’? (cf. Kretzenbacher et al. 2006).
Thanks for your attention!

natalia.levshina@uni-leipzig.de

The slides are available at

www.natalialevshina.com/presentations.html