A pilot study of T/V-forms in European languages based on a parallel corpus of online film subtitles

Natalia Levshina
Leipzig University
Outline

1. Theoretical background: T/V-distinction
2. Data: film subtitles from ParTy corpus
3. Quantitative analyses:
   • Relative frequencies of T/V-forms
   • Communicative constraints: conditional inference trees and random forests
Object of study

• T/V-distinction in addressing the hearer/reader (only pronouns and verb forms)

• The distinction is present in most European languages
  • T-forms: informal, familiar, e.g. French tu, German du, Russian ты + Verb 2nd SG
  • V-forms: formal, polite, e.g. French vous, German Sie, Russian вы + Verb 2nd PL or 3rd SG/PL

• Go back to Latin tu and vos (addressing the Emperor in the plural)
Languages in the sample

• Germanic: German, Dutch, Swedish
• Romance: French, Spanish, Romanian
• Slavic: Russian, Polish, Bulgarian
• 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>3(^{rd}) person PL</td>
</tr>
<tr>
<td>Dutch</td>
<td>2</td>
<td>jij (je)</td>
<td>u</td>
<td>2(^{nd}) person SG</td>
</tr>
<tr>
<td>Swedish</td>
<td>2</td>
<td>du</td>
<td>ni</td>
<td>2(^{nd}) PL</td>
</tr>
<tr>
<td>French</td>
<td>2</td>
<td>tu</td>
<td>vous</td>
<td>2(^{nd}) PL</td>
</tr>
<tr>
<td>Spanish</td>
<td>2</td>
<td>tú</td>
<td>usted</td>
<td>3(^{rd}) person SG</td>
</tr>
<tr>
<td>Romanian</td>
<td>3</td>
<td>tu</td>
<td>dumneata, Dumneavoastră</td>
<td>2(^{nd}) SG/PL 2(^{nd}) PL</td>
</tr>
<tr>
<td>Russian</td>
<td>2</td>
<td>ты [ty]</td>
<td>вы [vy]</td>
<td>2(^{nd}) PL</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>2</td>
<td>ти [ti]</td>
<td>Вие ['vi.ε]</td>
<td>2(^{nd}) PL</td>
</tr>
<tr>
<td>Polish</td>
<td>2</td>
<td>ty</td>
<td>pan (m)/pani (f)</td>
<td>3(^{rd}) person SG</td>
</tr>
<tr>
<td>Greek</td>
<td>2</td>
<td>еσύ [e'si]</td>
<td>еσείς [e'sis]</td>
<td>2(^{nd}) PL</td>
</tr>
<tr>
<td>Finnish</td>
<td>2</td>
<td>sinä</td>
<td>te</td>
<td>2(^{nd}) PL</td>
</tr>
</tbody>
</table>
Cross-linguistic research: type-based

• WALS Chapter 45, Helmbrecht 2013
Token-based typology

• However, there are many other interesting questions that can be asked:
  • 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?

• Not much research, so far...
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.
  • Later adopted for communication within a social group: e.g. the 17th century French nobility and bourgeoisie always used V-forms when speaking to one another, while servants used T-forms between themselves.

• 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 subtle cross-linguistic differences
Outline

1. Theoretical background: T/V-distinction
2. Data: ParTy corpus
3. Quantitative analyses:
   • Relative frequencies of T/V-forms
   • Communicative constraints: conditional inference trees and random forests
ParTy corpus

• a Parallel corpus for Typology
• subtitles of films and TED talks
• mostly Indo-European languages, but also other major languages (Chinese, Turkish, Finnish, Indonesian, Japanese, Thai, etc.)
• all languages aligned with English
• downloadable files at www.natalialevshina.com/corpus.html
• work in progress...
Films

INCEPTION

AVATAR

THE IRON LADY

FROZEN

BLACK SWAN

GRAND BUDAPEST HOTEL
Data set

• English data: instances of *you/yourselves* used when referring to one person. The plural uses are disregarded.
• 158 communicative situations with unique participants (in order to ensure maximal diversity)
• Translations into 11 languages coded wrt. T- or V-forms used
Example

• 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)
Communicative variables

• Dyadic asymmetric (power):
  • Age: is the hearer younger, older or of the same age (approximately) than/as the speaker?
  • Power: does the hearer have social power over the speaker? E.g. employer wrt. employee, prime-minister wrt. minister, general wrt. soldier
  • Gender: M to M, M to F, F to M, F to F

• Dyadic symmetric (solidarity):
  • Circle: self > family > romance, friends > work relationships > acquaintances > strangers
    + “house” (household servants, hotel, prison)
Communicative variables (cont.)

• Individual:
  • Age of the speaker (child, adult, elderly)
  • Age of the hearer (child, adult, elderly)
  • Social class of the speaker (upper, middle, lower)
  • Social class of the hearer (upper, middle, lower)
  • Gender of the speaker
  • Gender of the hearer
Example

*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

• Dyadic asymmetric:
  • Age: the hearer is younger
  • Social power: lower
  • Gender: F to M

• Dyadic symmetric:
  • Circle: romance

• Individual:
  • Age of speaker: old
  • Age of hearer: middle
  • Class of speaker: upper
  • Class of hearer: lower
  • Gender of speaker: F
  • Gender of hearer: M
Outline

1. Theoretical background: T/V-distinction
2. Data: ParTy corpus
3. Quantitative analyses:
   • Relative frequencies of T/V-forms
   • Communicative constraints: conditional inference trees and random forests
Proportions of T/V-forms in the data set in 11 languages

Proportions of T-V forms

V-forms
T-forms
fre ger rus dut bul spa pol swe fin gre rom

[Bar chart showing proportions for each language]
Some notes

• The highest proportions of V-forms are found in the languages with obligatory subject pronouns
  • except Swedish, where the spread of the T-pronoun was a political issue in the 1960s

• A hypothesis: Is this because the use of T-verb forms without pronouns may be perceived as less face-threatening than the use of T-verb forms AND T-pronouns?
Outline

1. Theoretical background: T/V-distinction
2. Data: ParTy corpus
3. Quantitative analyses:
   • Relative frequencies of T/V-forms
   • Communicative constraints: conditional inference trees and random forests
Conditional inference trees and random forests

- Predict the choice between the T- and V-forms in each language based on the communicative variables
- Robust in situations of strongly correlated predictors and in situations of many predictors and few observations
- Forests are grown from many conditional trees
Conditional inference tree: German data
Variable importance: German

Based on 500 trees, mtry = 5
Across the trees and forests...

• A substantial part of variation is due to the situations shown in particular films and/or strategies chosen by individual translators

• Circle has some importance in all languages -> solidarity dimension

• Some influence of Social Class, but only individually, not asymmetrically

• Exceptions:
  • Finnish and Swedish, where Gender is prominent
  • Greek: Gender and Age (but only a weak tendency)
Conditional inference tree: Finnish data
Greek data tree (alpha = 0.5):
Future research

• More data (whether the differences are significant)
• New variables: time (e.g. before and after 1960s), British and American
• Investigate the relationship between pro-drop and the relative frequencies of T/V-forms
Thank you!

The slides are available at
www.natalialevshina.com/presentations.html

Questions? Suggestions?
natalevs@gmail.com