Predicting Dialect Variation In Immigrant Contexts
Using Light Verb Constructions

A. Seza Doğruöz (NIAS) & Preslav Nakov (QCRI)

a.s.doigruoz@gmail.com & pnakov@qf.org.qa

Summary

**Motivation:**
- Immigrant languages (e.g., Turkish) change due to contact with the language (e.g., Dutch) of the host communities.
- Most NLP tools ignore immigrant varieties (NL-Turkish) and focus on the standard ones (TR-Turkish).
- Long-term goal:
  - Translations that take immigrant varieties into account.
  - Dialect-aware NLP tools

**Our Objective:** Distinguish between
- Turkish spoken in the Netherlands (NL-Turkish) vs.
- Turkish spoken in Turkey (TR-Turkish).

**Focus:** Light Verb Constructions

Contributions

- We are the first to predict on-going dialect variation in immigrant contexts as opposed to studying established dialect variations.
- We are also the first to compare bilingual LVCs with monolingual ones across two dialects of the same language.
- Our comparison of grammatical versus contextual features reveals context to be much more important.
- We experiment with LVCs extracted from natural spoken data rather than relying on isolated occurrences out of context.

Nominal Features

1. Etymological Origin
   - Dutch

   Example (1)
   
   *0 arkaadaş overplaat-sen yap-ı-aca-t-s.*
   That friend replace-inf do-pass-fut-past
   "That friend would have been replaced"

   Arabic Influence

   Example (2)
   
   Hoca-m diye
   Teacher-poss.1sg. as
   hitap ed-i-yo-z biz.
   address do-prog-1pl we
   "We address (him) as the teacher."

2. Case Marking (presence or absence)

   Example (3)
   
   Bazen yemek yap-ar-di-m.
   Sometimes food do-pres-past-1sg.
   "I used to prepare food sometimes."

Verbal Features

1. Finiteness

   Example (4)
   
   *Mısafir-ler-e ikram et-mek için al-di-k.*
   Guest-pl-dat serve do-inf for buy-past-1pl
   "We bought it to serve the guests."

2. Type of the Light Verb

   Example (5)
   
   Orda kadın doğum et-ti. (instead of "yap")
   There lady birth do-past.
   "The lady gave birth there."

3. Word Order in LVCs (OV or VO)

   Example (6)
   
   Yap-acak bir şey yok.
   Do-part one thing exist.not.
   "There is nothing to do."

   Context

   We also included the words surrounding the LVCs.

Results

<table>
<thead>
<tr>
<th>Features</th>
<th>Left vs. Right</th>
<th>No Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>56.38</td>
</tr>
<tr>
<td>Full Model</td>
<td>82.81</td>
<td>84.30</td>
</tr>
<tr>
<td>No context</td>
<td>70.67</td>
<td></td>
</tr>
<tr>
<td>No nominal complements</td>
<td></td>
<td>82.19</td>
</tr>
<tr>
<td>No info about etymological origin</td>
<td></td>
<td>82.10</td>
</tr>
<tr>
<td>No finiteness</td>
<td>83.03</td>
<td>84.35</td>
</tr>
<tr>
<td>No case marking</td>
<td>82.76</td>
<td>84.43</td>
</tr>
<tr>
<td>No word order</td>
<td>82.89</td>
<td>84.43</td>
</tr>
<tr>
<td>No verb type</td>
<td>82.94</td>
<td>84.39</td>
</tr>
</tbody>
</table>

Data

- NL-Turkish spoken corpus: 46 speakers from the Netherlands (74,461 words)
- TR-Turkish spoken corpus: 22 speakers from Turkey (28,731 words)

Corpora #etmek #yapmak #Total
NL-Turkish  449  543  992
TR-Turkish  527  755  1282
Total       976  1298

Experiments

Features: (1) words from the LVC context, (2) type of the light verb (yapmak or etmek), (3) the nominal complements, (4) finiteness of the verb (finite/non-finite), (5) case marking on the nominal complement (yes/no), (6) word order (VO/OV), (7) etymological origins of the nominal complement (Arabic/Dutch/French/English/Persian/Turkish/mixed).

Classifier: SVM with linear kernel

Evaluation: 5-fold cross validation

Two experiments: distinguish left/right context?

Discussion

Most important feature: Context

Also important:
- nominal complements;
- etymological origin.

Future Work

- other dialects in immigrant settings (e.g., Turkish spoken in Germany);
- other MWEs (e.g., noun compounds).