Predicting Dialect Variation in Immigrant Contexts Using Light Verb Constructions

A. Seza Doğruöz
Netherlands Institute for Advanced Study
Wassenaar, Netherlands
a.s.dogruoz@gmail.com

Preslav Nakov
Qatar Computing Research Institute
Tornado Tower
floor 10, P.O. Box 5825, Doha, Qatar
pnakov@qf.org.qa

Abstract

Languages spoken by immigrants change due to contact with the local languages. Capturing these changes is problematic for current language technologies, which are typically developed for speakers of the standard dialect only. Even when dialectal variants are available for such technologies, we still need to predict which dialect is being used. In this study, we distinguish between the immigrant and the standard dialect of Turkish by focusing on Light Verb Constructions. We experiment with a number of grammatical and contextual features, achieving over 84% accuracy (56% baseline).

1 Introduction

Human languages are in constant evolution, driven in part by contact with other languages (Uriel, 1953; Thomason, 2008). In immigrant contexts, bilingual and multilingual speakers act as agents of change by transmitting borrowed words and expressions across languages (Grosjean, 2014). Depending on social factors such as duration and intensity of contact with the local languages, large-scale spread of borrowed elements could lead to differences between the contact and non-contact dialects of the same language (Winford, 2005). For example, Spanish spoken by immigrants in USA sounds different in comparison to Spanish spoken in South America (Corvalán, 2003).

In this study, we focus on the immigrant dialect of Turkish as spoken in the Netherlands (NL-Turkish), which differs from Turkish spoken in Turkey (TR-Turkish). In contact situations, it is common for verbs to be borrowed across languages and integrated as nominal complements of Light Verb Constructions (LVCs) (Edwards and Gardner-Chloros, 2007; Butt, 2010). NL-Turkish LVCs are changing due to Dutch influence (Doğruöz and Backus, 2007; Doğruöz and Backus, 2009; Doğruöz and Gries, 2012). However, assessing Dutch influence is not always easy since NL-Turkish LVCs still co-exist with the TR-Turkish LVCs. This study aims to automatically identify the features that can distinguish between NL-Turkish and TR-Turkish LVCs.

Our study would benefit Machine Translation systems targeting dialectal variation. It differs from studies concerning the well-established dialectal variations of Arabic, e.g., Levantine, Gulf, Egyptian, Maghrebi (Salloum and Habash, 2012), EU vs. Brazilian Portuguese (Marujo et al., 2011) or Turkish vs. Tatar (Altintas and Cicekli, 2002). In contrast, we are interested in developing language technologies for *immigrant* dialects, which are often understudied and lack written resources due to their unofficial status. When immigrant speakers face communication difficulties (e.g., bureaucratic affairs with the local officials, teacher-parent meetings, doctor-patient conversations) in the local languages (e.g., Dutch) of the host country, they are often provided with translation equivalents in the standard dialect (e.g., TR-Turkish) of their native languages. However, these translations ignore the evolution of the immigrant dialect.[1] By identifying the differences between two dialects of the same variety, we aim to improve Machine Translation systems targeting immigrant speakers. Our contributions are the following:

- 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 the monolingual ones across two dialects of the same language.

[1] One of the authors failed the driving test in the Netherlands due to the dialect variation in the Turkish translation.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1391–1395,
October 25-29, 2014, Doha, Qatar.

- 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 occurences, out of context.

2 Method

We follow Baldwin and Kim (2010) and Butt (2010) in their definitions of LVCs, which state that there is a unity between the nominal and the verbal complements, but the meaning of the verb is somewhat bleached. In this study, we focus on Turkish LVCs with the verbal complements of *yapmak*/*etmek*, which both can be translated as "make/do". LVCs with these verbal complements are undergoing change in NL-Turkish (Doğruöz and Backus, 2009).

We experiment with the following features to predict NL-Turkish vs. TR-Turkish LVCs.

2.1 Nominal Features

In addition to traditional LVCs (e.g. [*ütü yapmak*] "iron do" (to iron) with both complements of Turkish origins), there is also foreign influence on Turkish LVCs. Section 2.1.1 describes the foreign influence on both NL-Turkish and TR-Turkish nominal complements based on their etymological origins.

2.1.1 Influence on Nominal Complements

Dutch Influence In example (1), the Dutch verb *overplaats* is nominalized through the infinitive marker (*-en*) and precedes the Turkish verb *yapmak* to form a Turkish-Dutch bilingual LVC.

Example 1:
O arkadaş [**overplaats-en yap-ıl-acak-tı**.]
That friend [replace-inf[2] do-pass-fut-past].
That friend would have been replaced.

In addition to borrowing nominalized Dutch verbs to form bilingual LVCs, Dutch LVCs are also translated as a chunk into NL-Turkish. These translated LVCs sound unconventional to TR-Turkish speakers (Doğruöz and Gries, 2012). In example (2), the LVC [*sınav yapmak*] "exam do" is a literal translation of the Dutch [*examen doen*] "exam-pl do", which is used to describe how students take high school exams to graduate.

[2]acc: accusative, fut:future, inf:infinitive, past:past tense, part: participle, pres: present tense, pl: plural, poss: possesive, prog:progressive tense, sg: singular

In a similar context, TR-Turkish speakers would have used [*sınav-a girmek*] "exam enter" instead. These LVCs are also classified as having their origins in another language.

Example 2:
Üç gündür [**sınav yap-ıyor-uz**].
Three day [exam do-prog-1pl].
We are having exams for the last three days.

Other Foreign Influences Although Dutch influence is clearly present in NL-Turkish LVCs, TR-Turkish LVCs are also not free of foreign influence. We have come across Arabic, Persian, French and English influences on Turkish LVCs with nominalized foreign verbs or literally translated LVCs as chunks. Example (3) illustrates how a borrowed Arabic verb (*hitap*, "address") is integrated as a nominal complement into a Turkish LVC [*hitap etmek*] "address do".

Example 3:
Hoca-m diye [**hitap edi-yo-z**] biz.
Teacher-poss.1sg like [address do-prog-1pl] we.
We address (him) as the teacher.

Example (4) illustrates how an English LVC [*do sports*] is borrowed into Turkish as a chunk [*spor yapmak*] "sports do".

Example 4:
Yazın [**spor yap-ıyo-z**].
summer spor do-prog-1pl
We do sports in summer.

We have identified the etymological origins of LVCs in both corpora using an online etymological dictionary.[3] Although LVCs of Dutch origin only occur in NL-Turkish, LVCs borrowed from other languages (e.g., Arabic, English, French) occur both in NL-Turkish and in TR-Turkish.

2.1.2 Case Marking

We also came across Turkish [N V] constructions with "*yapmak*" and "*etmek*" where the nominal complement acts as the object of the verb.

Turkish marks the direct objects with accusative case marking if they are definite (Enç, 1991). In example (5), the nominal element is the object of the verb, and thus it has the accusative marker.

Example 5:
Ben kendi [**iş-im-i yap-ıyor-um.**]
I own [work-poss.1sg-acc do-prog-1sg].
I do my own work.

[3]http://www.nisanyansozluk.com/

However, indefinite objects of the verb are left unmarked for case. In example (6), *yapmak* takes an indefinite object (*food*) as the complement. The boundary between [N V] constructions with indefinite nominal objects and LVCs are somewhat blurry. In both cases, the meaning of the verbal complement is bleached out and the nominal complement weighs heavier than the verbal one. We will not dwell further on this subtle distinction, but we plan future work on this topic following Cook et al. (2007) and Vincze et al. (2013).

Example 6:
Bazen [**yemek yap-ar-dı-m**]
Sometimes [food do-pres-past-1sg]
I used to sometimes prepare food.

Since Dutch does not mark objects of the verb morphologically, NL-Turkish speakers have difficulty (e.g., unnecessary addition or omission of case markers) in determining the definiteness of the nominal complements in [N V] constructions (Doğruöz and Backus, 2009). Therefore, we expect this feature to differentiate well between NL-Turkish and TR-Turkish [N V] constructions and LVCs with *yapmak/etmek* as verbal complements.

2.2 Verbal Complements

2.2.1 Finiteness

The verbs in LVCs are assumed to be flexible for inflection (Baldwin and Kim, 2010). However, we know little about how fineteness contributes to the formation of LVCs. To the best of our knowledge, finiteness has not been tested as a feature for identifying LVCs earlier. Therefore, we encoded the finiteness on *yapmak/etmek* as a binary (yes/no) feature in both data sets. Example (7) illustrates a non-finite LVC where the verb stem (*et*) is accompanied with an infinitive marker (*-mek*).

Example 7:
Misafir-ler-e [**ikram et-mek**] için al-dı-k
Guest-pl-dat [serve do-inf.] for buy-past-1pl
We bought (it) to serve the guests.

2.2.2 Type

NL-Turkish speakers could use other light verbs than TR-Turkish speakers for the same LVC construction. In example (8), the NL-Turkish speaker uses [*doğum etmek*] "birth do" instead of [*doğum yapmak*] "birth do", which is commonly preferred by TR-Turkish speakers. To capture this difference between the two dialects, we include the verb type as a feature as well.

Example 8:
Orda kadın [**doğum et-ti**].
There lady [birth do-past].
The lady gave birth there.

2.3 Word Order in LVCs

To the best of our knowledge, the influence of word order in LVCs has not been investigated as a feature. Although Turkish has a relatively flexible constituent order, object-verb (OV) is the most frequent word order for both NL-Turkish and TR-Turkish (Doğruöz and Backus, 2007). NL-Turkish speakers have adopted Dutch word order verb-object (VO) for some syntactic constructions, but we know little about the word order variation for LVCs. Encoding the word order of LVCs as a binary feature (OV vs. VO) could give us clues about differences or similarities of LVC use in NL-Turkish and in TR-Turkish. In example (9), the nominal complement (*one thing*) follows the verbal complement instead of preceding it as seen in earlier examples.

Example 9:
[**Yap-acak bir şey**] yok.
[Do-part. one thing] exist.not
There is nothing to do.

2.4 Context

So far, most studies were carried out ignoring the context of LVCs but focusing on their inherent grammatical features (e.g., lexical, syntactic, semantic or morphological). However, the context of an utterance could potentially provide additional useful cues. Since our data comes from natural conversations, we also experimented with the contextual information (words surrounding LVCs) as a feature for both data sets.

3 Data

Our data comes from spoken NL-Turkish (46 speakers from the Netherlands, 74,461 words) and TR-Turkish (22 speakers from Turkey, 28,731 words) corpora collected by one of the authors. LVC's are automatically extracted from the data using their stem forms ("yap-", "et-" without the infinitive -mEk). Table 1 illustrates the frequency of [N V] constructions with *etmek* and *yapmak* in both data sets.

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

Table 1: Distribution of *etmek* and *yapmak*.

4 Experiments

Our aim is to build a classifier that can determine whether a particular utterance containing an LVC (with the verbs *yapmak* or *etmek*) is uttered by an NL-Turkish or a TR-Turkish speaker.

We make use the following features in our classifier: (1) words from the context of the LVCs, (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), and (7) etymological origins of the nominal complement (Arabic/Dutch/French/English/Persian/Turkish/mixed).

For the contextual features, we experiment with two models: (a) we distinguish between a word extracted from the context to the left or to the right of the verb (*yapmak* or *etmek*) in the feature space, and (b) we do not make a distinction in terms of context. The reason to experiment with option (a) is due to the potential importance of the word order. While the word order variation is already modeled through feature (6), we also include the context as an additional feature to test its effect. On the down side, adding context doubles the feature space size and could lead to data sparseness issues. For the context words, we did not filter out stopwords since they are part of natural speech.

For our experiments, we used an SVM classifier as implemented in LibSVM. We used a linear kernel; more complex kernels did not help. We report results for a 5-fold cross-validation.

5 Results

Table 2 illustrates the results of our experiments. All models outperform the majority class baseline of always predicting TR-Turkish (which is 56.38% accuracy) by a sizable margin. Furthermore, splitting the context into left/right yields approximately 1.5% absolute drop in accuracy.

	Split the Context?	
Features	**Left vs. Right**	**No Split**
Baseline	56.38	
Full model	**82.81**	**84.30**
no context	70.67	
no nominal complements	82.19	83.64
no info about etymol. origin	82.10	83.99
no finiteness	83.03	84.35
no case marking info	82.76	84.43
no word order info	82.89	84.43
no verb type	82.94	84.39

Table 2: Cross-validation accuracy (5 folds).

The lower part of the table shows the results when turning off each of the feature types. The context seems to be the most important feature since its exclusion leads to a drop from low-to-mid eighties to about 70% accuracy. Except the nominal complements and the information about etymological origins, most other features seem to have marginal impact on accuracy. Excluding the two features (nominal complements and etymological origins) lead to approximately 0.5% absolute drop in accuracy. The impact of the last four features in the table is tiny; excluding some of them even leads to a tiny improvement.

Overall, we can conclude that by far the most important features are the context features (without the left/right context split). The other useful features are the nominal complements and the information about the etymological origin of the borrowed LVCs. The remaining four linguistic features seem to be largely irrelevant.

6 Conclusion and Future Work

Language technologies are usually developed for standard dialects, ignoring the linguistic differences in other dialects such as those in immigrant contexts. One of the reasons for this is the difficulty of assessing and predicting linguistic differences across dialects. This is similar to efforts to translate well-established Arabic dialects (Bakr et al., 2008; Sawaf, 2010), or to adapt between Brazilian and European Portuguese (Marujo et al., 2011), Czech–Slovak (Hajič et al., 2000), Spanish–Portuguese (Nakov and Ng, 2009; Nakov and Ng, 2012), Turkish–Crimean Tatar (Altintas and Cicekli, 2002), Irish–Scottish Gaelic (Scannell, 2006), Bulgarian–Macedonian (Nakov and Tiedemann, 2012), Malay–Indonesian (Wang et al., 2012) or Mandarin–Cantonese (Zhang, 1998).

In this work, we have built a classifier that uses LVCs to differentiate between two different Turkish dialects: standard and immigrant. The results indicate that contextual features are most useful for this task. Although this requires further investigation, we can explain it by the thousands of features context generates: each contextual word is a feature. Thus, it is very hard for our grammatical features to compete against contextual features but they do have an impact.

We are planning to extend our study to dialects in other immigrant settings (e.g., Turkish in Germany) and to other types of multiword expressions (e.g., [N N] compounds).

References

Kemal Altintas and Ilyas Cicekli. 2002. A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences*, ISCIS '02, pages 192–196, Orlando, FL, USA.

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *Proceedings of the 6th International Conference on Informatics and Systems*, INFOS '08, Cairo, Egypt.

Timothy Baldwin and Su Nam Kim, 2010. *In Nitin Indurkhya and Fred J. Damerau (eds.), Handbook of Natural Language Processing*, chapter Multiword expressions, pages 267–292. CRC Press, Boca Raton, USA, second edition.

Miriam Butt, 2010. *In Mengistu Amberber, Brett Baker, and Mark Harvey (eds.), Complex predicates: cross-linguistic perspectives on event structure*, chapter The light verb jungle: still hacking away, pages 48–78. Cambridge University Press.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of MWE '07*, pages 41–48, Prague, Czech Republic.

Carmen Silva Corvalán. 2003. Otra mirada a la expresión del sujeto como variable sintáctica. *Lengua, variación y contexto: Estudios dedicados a Humberto López Morales*, 2:849–860.

A. Seza Doğruöz and Ad Backus. 2007. Postverbal elements in immigrant Turkish: Evidence of change? *International Journal of Bilingualism*, 11(2):185–220.

A. Seza Doğruöz and Ad Backus. 2009. Innovative constructions in Dutch Turkish: An assessment of ongoing contact-induced change. *Bilingualism: Language and Cognition*, 12(01):41–63.

A. Seza Doğruöz and Stefan Gries. 2012. Spread of on-going changes in an immigrant language: Turkish in the Netherlands. *Review of Cognitive Linguistics*, 10(2).

Malcolm Edwards and Penelope Gardner-Chloros. 2007. Compound verbs in codeswitching: Bilinguals making do? *International Journal of Bilingualism*, 11(1):73–91.

Mürvet Enç. 1991. The semantics of specificity. *Linguistic Inquiry*, 22(1):1–25.

François Grosjean. 2014. Bicultural bilinguals. *International Journal of Bilingualism*, pages 1–15.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of ANLP '00*, pages 7–12, Seattle, WA, USA.

Luís Marujo, Nuno Grazina, Tiago Luís, Wang Ling, Luísa Coheur, and Isabel Trancoso. 2011. BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of EAMT '11*, pages 129–136, Leuven, Belgium.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of EMNLP '09*, pages 1358–1367, Singapore.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *J. Artif. Intell. Res. (JAIR)*, 44:179–222.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of ACL '12*, Jeju Island, Korea.

Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to standard Arabic machine translation system. In *Proceedings of COLING '12*, pages 385–392, Mumbai, India.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of AMTA '10*, Denver, Colorado.

Kevin Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the LREC 2006 Workshop on Strategies for developing machine translation for minority languages*, pages 103–107, Genoa, Italy.

Sarah Thomason. 2008. Social and linguistic factors as predictors of contact-induced change. *Journal of language contact*, 2(1):42–56.

Weinreich Uriel. 1953. Languages in contact: Findings and problems. *Publications of the Linguistic Circle of New York*, vol. 1.

Veronika Vincze, István Nagy, and Richárd Farkas. 2013. Identifying English and Hungarian light verb constructions: A contrastive approach. In *Proceedings of ACL '13*, pages 255–261, Sofia, Bulgaria.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2012. Source language adaptation for resource-poor machine translation. In *Proceedings of EMNLP-CoNLL '12*, pages 286–296, Jeju Island, Korea.

Donald Winford. 2005. Contact-induced changes: Classification and processes. *Diachronica*, 22(2):373–427.

Xiaoheng Zhang. 1998. Dialect MT: a case study between Cantonese and Mandarin. In *Proceedings of the COLING '98*, pages 1460–1464, Montreal, Quebec, Canada.