# First Workshop on Language Resources and Technologies for Turkic Languages

## Workshop Programme

14:00 – 14:10 Welcome
14:10 – 15:10 Oral Session - I

- Cengiz Acartürk and Murat Perit Çakır, *Towards Building a Corpus of Turkish Referring Expressions*
- Arianna Bisazza and Roberto Gretter, *Building a Turkish ASR System with Minimal Resources*
- Francis Tyers, Jonathan North Washington, Ilnar Salimzyanov and Rustam Batalov, *A Prototype Machine Translation System for Tatar and Bashkir Based on Free/Open-Source Components*

15:10 – 15:30 Poster Presentations

- Işın Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban, Ruket Çakıcı and Deniz Zeyrek, *Turkish Discourse Bank: Ongoing Developments*
- Seza Doğruöz, *Analyzing Language Change in Syntax and Multiword Expressions: A Case Study of Turkish Spoken in the Netherlands*
- Atakan Kurt and Esma Fatma Bilgin, *The Outline of an Ottoman-to-Turkish Machine Transliteration System*
- Vít Baisa and Vít Suchomel, *Large Corpora For Turkic Languages and Unsupervised Morphological Analysis*
- Ayışığı B. Sevdik-Çallı, *Demonstrative Anaphora in Turkish: A Corpus Based Analysis*
- Alexandra V. Sheymovich and Anna V. Dybo, *Towards a Morphological Annotation of the Khakass Corpus*

15:30 – 16:30 Coffee Break & Poster Session

16:30 – 17:50 Oral Session - II

- Benjamin Mericli and Michael Bloodgood, *Annotating Cognates and Etymological Origin in Turkic Languages*
- Özkan Kılıç and Cem Bozşahin, *Semi-Supervised Morpheme Segmentation without Morphological Analysis*
- Şükriye Ruhi, Kerem Eryılmaz and M. Güneş C. Acar, *A Platform for Creating Multimodal and Multilingual Spoken Corpora for Turkic Languages: Insights from the Spoken Turkish Corpus*
- Eray Yıldız and A. Cüneyd Tantuğ, *Evaluation of Sentence Alignment Methods for English-Turkish Parallel Texts*

17:50 – 18:00 Closing

## Editors

| | |
|---|---|
| Şeniz Demir | Tübitak-Bilgem |
| İlknur Durgar El-Kahlout | Tübitak-Bilgem |
| Mehmet Uğur Doğan | Tübitak-Bilgem |

## Workshop Organizers/Organizing Committee

| | |
|---|---|
| Kemal Oflazer | Carnegie Mellon University - Qatar |
| Mehmed Özkan | Boğaziçi University |
| Mehmet Uğur Doğan | Tübitak-Bilgem |
| Hakan Erdoğan | Sabancı University |
| Dilek Hakkani-Tür | Microsoft |
| Yücel Bicil | Tübitak-Bilgem |
| İlknur Durgar El-Kahlout | Tübitak-Bilgem |
| Şeniz Demir | Tübitak-Bilgem |
| Alper Kanak | Tübitak-Bilgem |

## Workshop Programme Committee

| | |
|---|---|
| Yeşim Aksan | Mersin University |
| Adil Alpkoçak | Dokuz Eylül University |
| Mehmet Fatih Amasyalı | Yıldız Technical University |
| Ebru Arısoy | IBM T.J. Watson Research Center |
| Levent Arslan | Boğaziçi University |
| Barış Bozkurt | Bahçeşehir University |
| Cem Bozşahin | Middle East Technical University |
| Ruket Çakıcı | Middle East Technical University |
| Özlem Çetinoğlu | University of Stuttgart |
| Cemil Demir | Tübitak-Bilgem |
| Cenk Demiroğlu | Özyeğin University |
| Banu Diri | Yıldız Technical University |
| Gülşen Cebiroğlu Eryiğit | İstanbul Technical University |
| Engin Erzin | Koç University |
| Tunga Güngör | Boğaziçi University |
| Ümit Güz | Işık University |
| Yusuf Ziya Işık | Tübitak-Bilgem |
| Selçuk Köprü | Teknoloji Yazılımevi |
| Atakan Kurt | Fatih University |
| Oğuzhan Külekçi | Tübitak-Bilgem |
| Coşkun Mermer | Tübitak-Bilgem |
| Arzucan Özgür | Boğaziçi University |
| Fatma Canan Pembe | Tübitak-Bilgem |
| Şükriye Ruhi | Middle East Technical University |
| Murat Saraçlar | Google - Boğaziçi University |
| Bilge Say | Middle East Technical University |
| Ahmet Cüneyd Tantuğ | İstanbul Technical University |
| Erdem Ünal | Tübitak-Bilgem |
| Deniz Yüret | Koç University |
| Deniz Zeyrek | Middle East Technical University |

# Table of Contents

# Author Index

# Introduction

Turkic languages are spoken as a native language by more than 150 million people all around the world (one of the 15 most widely spoken first languages). Prominent members of this family are Turkish, Azerbaijani, Turkmen, Kazakh, Uzbek, and Kyrgyz. Turkic languages have complex agglutinative morphology with very productive inflectional and derivational processes leading to a very large vocabulary size. They also have a very free constituent order with almost no formal constraints. Furthermore, due to various historical and social reasons these languages have employed a wide-variety of writing systems and still do so. These aspects bring numerous challenges (e.g., data sparseness and high number of out-of-vocabulary words) to computational processing of these languages in tasks such as language modeling, parsing, statistical machine translation, speech-to-speech translation, etc. Thus, pursuing high-quality research in this language family is particularly challenging and laborious.

This workshop is timely as there is burgeoning interest in the field of research. Moreover, various language resources and computational processing techniques for Turkic languages need to be developed in order to bring their status up to par with more studied languages in the context of speech and language processing. It has become more crucial as the number of international affairs, economic activities, and cultural relations between Turkic people and EMEA (Europe, Middle East, and Africa) increase. There exist a growing demand and awareness on related research and current developments provide us with solutions from different approaches. However, there still remain many problems to be solved and much work to be done in the roadmap for Turkic languages.

The workshop will bring together the academicians, experts, research-oriented enterprises (SMEs, large companies, and potential end users), and all other stakeholders who are actively involved in the field of speech and language technologies for Turkic languages. The workshop will focus on cut-edge research and promote discussions to better disseminate knowledge and visionary thoughts for speech and language technologies aligned with Turkic languages. The workshop is expected to properly portray the current status of Turkic speech and language research performances, and to enlighten the pros and cons, end user needs, current state-of-the-art, and existing R&D policies and trend. This workshop will also have a positive impact on establishing a research community moving into the future and on building a collaboration environment which we anticipate to receive widespread attention in the HLT domain.

The workshop features 7 oral and 6 poster presentations. The accepted papers range from annotation initiatives to language and speech resources and technologies.

# Analyzing language change in syntax and multiword expressions: A case study of Turkish Spoken in the Netherlands

**A. Seza Doğruöz**
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
E-mail: a.s.dogruoz@gmail.com

### Abstract

All languages change and spoken corpora provide opportunities to analyze linguistic changes while they are still taking place. Turkish spoken in the Netherlands (NL-Turkish) has been in contact with Dutch for over fifty years and it sounds different in comparison to Turkish spoken in Turkey (TR-Turkish). Comparative analyses of NL-Turkish and TR-Turkish spoken corpora do not reveal significant on-going changes in terms of word order. However, Dutch-like multiword expressions make NL-Turkish sound unconventional to TR-Turkish speakers. In addition to presenting these on-going changes, this study also discusses the challenges with respect to syntactic parsing as well as identification and classification of multiword expressions in spoken Turkish corpora.

**Keywords:** multiword expressions, language variation and change, spoken corpus

## 1. Contact-Induced Language Change

What all languages share is changeability and contact with other languages is one of the reasons for change (Heine &Kuteva, 2005; Thomason, 2001; Weinreich, 1953). Language change is a gradual process with synchronic and diachronic aspects. The synchronic aspect (variation) refers to the occurrence of unconventional variants (i.e. innovations) at a given time in an utterance. The diachronic aspect (change), on the other hand, refers to the accumulation of these unconventional variants over time (Labov, 2010a; Labov, 2010b).

Explaining unconventional forms in a language start with finding their source. Generally, two main sources are distinguished: internal and external ones (Winford, 2003; Elsik & Matras, 2006). In the internal case, the source of the unconventionality is found within the language such as gradual changes (e.g. form, sound) over long periods of time. In the case of an external source, the unconventional form is copied from another language. This research focuses on Turkish-Dutch contact in the Netherlands where Dutch is the model language and serves as the source of change and Turkish is the replica language and undergoes change through Dutch influence. Turkish spoken in the Netherlands (NL-Turkish) sounds different in comparison to Turkish spoken in Turkey (TR-Turkish). Comparing NL-Turkish and TR-Turkish spoken corpora, this study investigates the on-going linguistic changes in NL-Turkish. More specifically, challenges with respect to syntactic parsing and identification of unconventional multiword expressions will be addressed.

## 2. How to identify structural changes: Word Order

Synchronically, there are two possibilities in producing an utterance (Croft, 2000:29):

- we comply with the conventions of the speech community we belong to and produce conventional forms
- we do not comply with the existing conventions and produce an unconventional (innovative) form.

Change only starts when an unconventional form is adopted by other members of the speech community.

One of the mechanisms through which structural innovations are introduced is the use of foreign morphemes and words (Weinreich, 1953; Thomason & Kaufman, 1988; Myers-Scotton, 2002). This is called code-switching and has been observed frequently in Turkish-Dutch contact (Boeschoten, 1990; Backus, 1996).

Languages borrow not only morphemes and words from each other but also grammatical relations such as structures (Johansson, 2002; Heine & Kuteva, 2005, Ross, 2007). One of those borrowed structures in contact situations is word order (Thomason, 2001; Heine, 2006). In Turkish-Dutch contact, the expectation is that Turkish (a Subject-Object-Verb language) will increase its SVO (Subject-Verb-Object) order due to contact with Dutch (a Subject-Verb-Object language). In order to test this claim, the relative frequencies of different word orders need to be measured and compared in the contact (NL-Turkish) vs. non-contact (TR-Turkish) varieties of Turkish. For example, if the SVO in NL-Turkish is relatively more frequent than the SVO in TR-Turkish, it is possible to say that NL-Turkish is undergoing change (probably) due to Dutch influence.

## 3. Method-I

This study makes use of NL-Turkish and TR-Turkish spoken corpora which were collected in the Netherlands and in Turkey respectively (Doğruöz, 2007). Transcribed part of NL-Turkish corpus measures about 328.000 words and TR-Turkish corpus measures about 170.000 words.

To my knowledge, there is currently no syntactic parser available for Turkish. Therefore, it is not possible to automatically assign syntactic roles in neither NL-Turkish nor TR-Turkish corpus. Using CLAN (Computerized Language Analysis) program, sample data sets in both NL-Turkish (24.200) words) and TR-Turkish corpora (20.210) were manually coded for syntactic roles in simplex clauses which include one finite verb (Doğruöz

& Backus, 2007). Example (1) illustrates how the coding was done.

(1)

| | | | |
|---|---|---|---|
| Anne-m | Oya-ya | oyuncak | al-dı. |
| Mother-POSS.1sg | Oya-DAT | toy | buy-PAST. |
| **S** | **IO** | **DO** | **V** |

## 4. Interim Results-I

The comparison of NL-Turkish and TR-Turkish corpora did not reveal any statistically significant differences in terms of (S)OV and (S)VO word orders (Doğruöz & Backus, 2007). However, (S)OV and (S)VO are attested as the most frequent and the least frequent word orders in both corpora respectively. This is in contrast with Gagauz, which is a Turkic language spoken in Moldova, Bulgaria and Ukraine for over 500 years (Menz, 1999). When the same manual coding system was applied to the Gagauz spoken conversations (based on transcripts provided in Menz, 1999), the results indicated that half of the simplex clauses had (S)VO word order (Doğruöz & Backus, 2007). In that sense, it is possible to claim that NL-Turkish may also change depending on the duration and intensity of contact with Dutch in the future. The availability of a syntactic parser could make it possible to compare word orders of Turkic languages with each other automatically and identify possible contact-induced effects in other Turkic languages as well.

## 5. How to identify structural changes: Multiword expressions

Frequency accounts are crucial for detecting the on-going structural changes but it is not always easy to know what to count. The reason is the difficulty of identifying the unit of the language that is targeted by a change. Typically, different structural levels of language are simultaneously involved in the production of an utterance.

One of the main issues in typological and cross-linguistic research is the difficulty of comparison since linguistic categories in one language may not correspond exactly to the categories in other languages. In other words, universal categories that would apply to each and every language are rarely existent (Evans & Levinson, 2009). Moreover, within a language, it is very difficult to establish sharp, clear-cut boundaries between different linguistic categories (Weinreich, 1953; Croft, 2007). Cognitive Linguistics provides a theoretical framework to identify multiword expressions since it does not recognize a traditional boundary between lexicon and syntax.

In daily life, we speak neither with isolated words (e.g. *drink*, *juice*) nor with highly abstract patterns (e.g. [V O]). Instead, we speak with highly fixed units [*good* evening] or partially schematic ones [*drink* NP] and produce full utterances (e.g. *Good evening, let's drink something*). What we encounter in daily life is not the abstract structures but rather specific instantiations of these structures. Based on our inventory of fixed and partially schematic multiword expressions we make generalizations and produce new utterances. Since language use and inventory depend on experience, these approaches are defined as "usage-based". Language is assumed to be made up of multiword expressions of different types and sizes and they have a unique form-meaning relationship in every language (Bybee, 2006).

This gradient view (Croft, 2007) fits very well with the phenomenon of language change since languages change in small steps. Although the analysis of NL-Turkish spoken corpus does not reveal sweeping syntactic changes in terms of word order, there are several multiword expressions that sound unconventional for TR-Turkish speakers (Doğruöz & Backus, 2009). Next section describes the method to identify and classify these unconventional multiword expressions.

## 6. Method-II

The following steps were followed to identify and analyze unconventional multiword expressions in a sample NL-Turkish corpus (23.061 words) (Doğruöz & Backus, 2009):

- All the multiword expressions that would sound unconventional to TR-Turkish speakers were identified manually.
- A panel of TR-Turkish judges were consulted in order to confirm or disconfirm the unconventionality in a particular multiword expression.
- A TR-Turkish equivalent for each NL-Turkish unconventional multiword expression was established in order to identify which linguistic aspect causes unconventionality.
- A sample TR-Turkish spoken corpus (27.057 words) was analyzed for the possible occurrences of unconventional multiword expressions.
- In order to detect Dutch influence, Dutch equivalents of the unconventional NL-Turkish multiword expressions were established through collaboration with native Dutch speakers.

## 7. Interim Results-II

After unconventional NL-Turkish multiword expressions are identified, they are classified based on what causes their unconventionality. The result of this exercise revealed two types of unconventional multiword expressions:

- *Lexically Fixed Multiword expressions*

NL-Turkish constructions contain additional or substituted lexical items in comparison to TR-Turkish equivalents due to literal translation from Dutch (Doğruöz & Backus, 2009). For example, the verb *okumak* "read" is sunstituted with *yapmak* "do" in example (2). The unconventionality in this case is not due to the borrowing of a single lexical item but rather due to the borrowing of a Dutch multiword expression as a whole (e.g. [*Fransızca yapmak*] "French do").

(2)

NL-TR: Okul-da iki sene İngilizce **yap-tı-m**.
School-loc two year English do-past-1sg
*"(I) did English for two years at school"*
TR-TR: Okul-da iki sene İngilizce oku-du-m.
School-loc two year English read-past-1sg
*"(I) read English at school for two years".*
NL: Ik heb twee jaar Engels gedaan op school.
I have two year English do-perf. at school
*"I did English for two years at school"*

- *Partially Schematic Multiword expressions*

These multiword expressions host both fixed (lexical and morphological) items and open slots (i.e. positions that host any element). For example, in [*Eat* NP], the verb "*eat*' is the lexically fixed item whereas [NP] could be filled with various other lexical items. In addition to borrowing lexically fixed multiword expressions, NL-Turkish speakers also borrow partially schematic multiword expressions. In example (3), the function word *bir* "one" is perceived as redundant by TR-Turkish speakers. In this case, NL-Turkish speaker literally translates the partially schematic [*een stuk of* Number N] "one piece of Number N" multiword expression from Dutch into Turkish (Doğruöz & Backus, 2009).

(3)
NL-TR: Burda **bir** on tane soru var-dır.
Here one ten piece question exist-pres.
*"There are (approx.) ten questions here."*
TR-TR: Burda on tane soru var-dır.
Here ten piece question exist-pres.
*"There are probably ten questions here."*
NL: Soms zijn er een stuk of tien vragen.
Sometimes are there one piece of ten questions
*"Sometimes there are (approx.) ten questions."*

Similarly, there are some on-going changes in NL-Turkish multiword expressions that include case marking on nominal lexical items. Transitive verbs usually mark direct objects with accusative case in Turkish. Since Dutch does not have case marking, NL-Turkish speakers sometimes delete or substitute the case marking in these multiword expressions. In example (4), the accusative marker in the [N-acc *sevmek*] "N-acc like" multiword expression is deleted probably due to the Dutch influence (Doğruöz & Backus, 2009).

(4)
NL-TR: Türk müziğ-i çok sev-iyor-um.
Turkish music-poss.3sg very like-prog-1sg
"*I like Turkish music a lot*"
TR-TR: Türk müziğ-i-ni çok sev-iyor-um.
Turkish music-poss.3sg-acc very like-prog-1sg
"*I like Turkish music a lot*"
NL: Ik houd van Turkse muziek.
I like of Turkish music.
*"I like Turkish music"*

Currently, both types of unconventional constructions are identified and classified manually. Although this is doable for a small sub-corpus, it is not feasible for larger corpora. Therefore, there is a need for developing a method in order to identify and parse these units automatically or semi-automatically.

## 8. Conclusion: What to do next?

Languages are not static and they change constantly. Spoken and written corpora provide us with the data to identify and analyze the on-going (synchronic) and completed changes (diachronic). This study focuses on synchronic language change through analyzing comparative spoken corpora in two varieties of Turkish (i.e. NL-Turkish vs. TR-Turkish). While doing these analyses, the following challenges are encountered:

In order to compare word orders across different varieties of Turkish (or Turkic languages), there is a need for a syntactic parser which could assign syntactic roles to the lexical items in utterances (for spoken corpora). One of the challenges for this parser would be to establish standard transcriptions across different spoken corpora. Secondly, a decision should be made with regard to which syntactic roles to assign.

The analyses of NL-Turkish corpus reveal that the on-going changes are currently taking place through lexically fixed and partially schematic multiword expressions. Although a sub-corpus could be analyzed manually to identify and classify these multiword expressions, automatic identification techniques are necessary to analyze larger corpora (also see Eryiğit, İlbay, Can, 2011).

Lexically specific multiword expressions are usually searchable by their key words in corpora. However, the open slots in partially schematic units and the agglutinative nature of Turkish (i.e. the fact that free and bound morphemes are attached to each other) provide challenges to search these units automatically in large corpora.

Despite the computational challenges presented above, spoken and written corpora provide excellent opportunities to uncover similar and different linguistic aspects across Turkic languages. In order to make these comparisons, there is a need for collaboration between the linguists who need to find answers to linguistic questions and computational linguists who will provide means to analyze the language data in different forms and shapes (Levin, 2011; Steedman, 2011).

## 9. References

Backus, A. (1996). *Two in one: Bilingual speech of Turkish immigrants in the Netherlands*. Tilburg: Tilburg University Press.

Boeschoten, H.E. (1990). *Acquisition of Turkish by immigrant children: A multiple case study of Turkish children in the Netherlands*. Wiesbaden: Harrowitz.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82, pp. 711-733.

Croft, W. (2000). *Explaining language change: an evolutionary approach.* Harlow, Essex: Longman

Croft, W. (2007). Beyond Aristotle and gradience: A reply to Aarts. Studies in Language, 31, pp. 409-430.

Doğruöz, A.S. (2007). Synchronic Variation and Diachronic Change in Dutch Turkish: A Corpus Based Analysis. *Ph.D. thesis. Tilburg University Press.*

Doğruöz, A.S., Backus, A. (2007). Postverbal elements in immigrant Turkish: Evidence of change? *International Journal of Bilingualism,* 11 (2), pp. 185-220.

Doğruöz, A. S., Backus, A. (2009). Innovative constructions in Dutch-Turkish: An assessment of on-going contact-induced change. *Bilingualism: Language and Cognition*, 12 (1), pp. 41-63.

Elsik, V., Matras, Y. (2006). *Markedness and Language Change: The Romani Sample*. Berlin: Mouton de Gruyter.

Eryiğit, G., Ilbay, T., Can, O.A. (2011). *Multiword Expressions in Statistical Dependency Parsing*. Proceedings of the 2.Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2011), 45-55.

Evans, N., Levinson, S.C. (2009). The myth of language universals: Language diversity and its importance for Cognitive Science. *Behavioral and Brain Sciences*, 32, pp. 429-492.

Heine, B., Kuteva, T. (2005). *Language contact and grammatical change*. Cambridge: Cambridge University Press.

Heine, B. (2006). Contact-induced word order change without word order change. Working papers in multilingualism. Series B, 76, Universitat Hamburg.

Johansson, L. (2002). *Structural factors in Turkic language contacts*. Richmond/Surrey: Curzon Press.

Labov, W. (2010a). *Principles of Linguistic Change, Volume I, Internal Factors*, Oxford: Blackwell.

Labov, W. (2010b). *Principle of Linguistic Change, Volume II, Social Factors*, Oxford: Blackwell.

Levin, L. (2011). Variety, Idiosyncrasy and complexity in Language and Language Technologies. *Linguistic Issues in Language Technology*, 6, pp. 1-22.

Menz, A. (1999). *Gagausische Syntax: Eine Studie zum kontakinduzierten Sprachwandel.* Wiesbaden: Harrosowitz.

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual encounters and grammatical outcomes*. New York: Oxfrd University Press.

Ross, M. (2007). Calquing and metatypy. *Journal of Language Contact*, THEMA 1, pp. 116-143.

Steedman, M. (2011). Romantics and Revolutionaries: What theoretical and computational linguists need to know about each other. *Linguistic Issues in Language Technology*, 6, pp.1-21.

Thomason, S.G. (2001). *Language Contact: An introduction.* Washington D.C.: Georgetown University Press.

Thomason, S.G., Kaufman, T. (1988). Language contact, creolization and genetic linguistics. Berkeley/Los Angeles: University of California Press.

Weinreich, U. (1953). *Languages in contact: Findings and problems.* The Hague: Mouton.

Winford, D. (2003). *An introduction to contact linguistics*. Oxford: Blackwell Publishing.