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.

Towards Building a Corpus of Turkish Referring Expressions

Cengiz Acartürk, Murat Perit Çakır
Department of Cognitive Science, Informatics Institute
Middle East Technical University
Universiteler Mah. 06800 Çankaya, Ankara, Turkey
E-mail: acarturk@metu.edu.tr, perit@metu.edu.tr

Abstract

In this paper we report on the preliminary findings of our ongoing study on Turkish referring expressions used in situated dialogs. Situated dialogs of pairs of Turkish speakers were collected while they were engaged with a collaborative Tangram puzzle solving task, which was designed by Spanger et al (2011) in an effort to build a corpus of referring expressions in Japanese and English. The paper provides our preliminary results on the Turkish corpus and compares them with the findings of comparable studies conducted on Japanese and English referring expressions.

Keywords: Referring expressions, multimodal corpora, discourse annotation, Turkish language resources

1. Introduction

Referring expressions are linguistic resources that allow speakers to identify objects relevant to their ongoing interaction. Reference production and understanding of references involve the ability to think of and represent objects, to direct others' attention to relevant objects in the shared scene, and to identify what other speakers are talking about when they use such expressions (Gundel & Heldberg, 2008). Therefore, referencing practices in which such expressions are put into use are essential for understanding how language mediates cognition at the intra and inter-subjective levels (Hanks, 2009).

Referring expressions have gained increased attention from computational linguists due to the interest towards developing more natural and efficient human-agent interactions in the real-world context. Recently several corpora have been created to aid the analysis of referring expressions in English. For instance, the COCONUT corpus (Di Eugenio et al., 2000) includes a repository of referring expressions used during text-based interactions in the context of a 2-D interior design task. QUAKE (Byron & Fosler-Lussier, 2006) and SCARE (Stoia et al., 2008) corpora are based on interactions recorded in the context of a collaborative treasure hunting task in a 3-D virtual world.

The work on these corpora has led to the development of useful categorization schemes for English referring expressions. However, due to the restrictions imposed on participants at each task scenario, these characterizations usually apply only to a subset of the rich variety of uses referring expressions may have in situated dialogs. For instance, the COCONUT task posed limitations on the use of language by restricting participants to use a text-based interface and enforcing a strict turn-taking protocol. The corpus also did not include extra-linguistic features relevant for understanding the use of referring expressions. In contrast, the QUAKE and SCARE corpora were collected in a voice-enabled 3-D world, which models a more complicated and realistic context of interaction. However, participants were restricted to carry out limited set of actions such as pushing buttons and picking up or dropping objects in this virtual world. For that reason, the QUAKE and SCARE corpora were mainly used for studying location-based references (Byron et al., 2005).

Referring expressions are particularly important in the context of collaborative activity where interlocutors need to establish a mutual orientation towards relevant objects in the scene to coordinate and make sense of each other's actions (Goodwin, 1996; Hanks; 1992; Clark & Wilkes-Gibbs, 1986). Existing corpora of referring expressions lack a naturalistic situated dialog context, which may have an influence on the type and distribution of referring expressions identified based on such corpora. This motivated Spanger et al. (2011) to design a collaborative problem solving activity where pairs of participants coordinate their actions with talk in an unrestricted way. Spanger et al.'s work led to the culmination of the REX-J corpus, which includes referring expressions in Japanese and English. This corpus differs from the previously discussed ones in terms of its focus on the study of event or action based references.

Relevant work on the Turkish language primarily focuses on pronoun disambiguation and anaphora resolution in text. A synchronic corpus of 2 million words (METU Corpus, MTC), a morphologically and syntactically tagged subcorpus of MTC with 65,000 words (Say et al., 2002), and a 500,000-word subcorpus of MTC with discourse annotation (Zeyrek et al., 2009) are recently available tagged corpora in Turkish. On the other hand, previous work relevant to the study of Turkish referring expressions involves pronoun disambiguation and anaphora resolution in text with natural language processing techniques (Kılıçaslan et al., 2009; Tin & Akman, 1994), and a computational model of

contextually appropriate anaphor and pronoun generation for Turkish (Yüksel & Bozşahin, 2002).

In this paper we appropriated Spanger et al.'s experimental setup in an effort to build a corpus of Turkish referring expressions. We aim to build and analyse a corpus that will guide subsequent work on a more general class of referring expressions used in *situated dialogs*. To the best of the authors' knowledge, our study is the first multimodal corpus that focuses on the use of referring expressions in situated, naturalistic dialogs in Turkish.

The rest of the paper is organized as follows. Section 2 provides an overview of the experimental setup and the annotation scheme used to build the Turkish referring expressions corpus. The next section summarizes the preliminary findings of our analysis on the Turkish corpus. The paper concludes with a comparison of our results with the findings of studies conducted on Japanese and English referring expressions, and with a discussion of some possible directions for future research.

2. Corpus Building

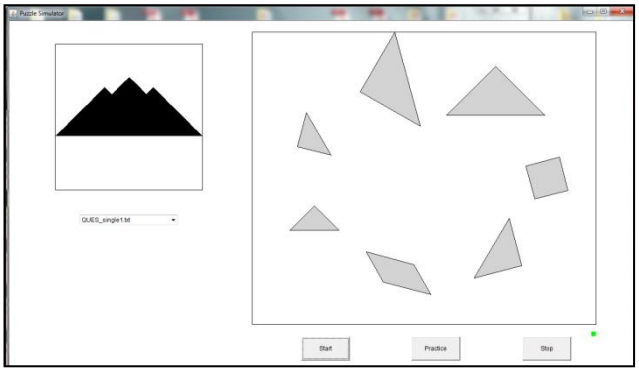

Figure 1: Screen shot of the Tangram Simulator software.

2.1 Experimental Setup

In the experiment, we employed the dual eye-tracking paradigm. Eight graduate students (2 female, 6 male) were recruited to participate in this study. The participants were grouped into 4 same-gender pairs. The pairs were located at different locations (two labs at METU campus) during the experiment. They coordinated their work through a screen sharing software called Team Weaver (www.teamweaver.com), which also enabled voice communication. Two non-intrusive eyetrackers (a Tobii T120 and a Tobii T1750) and the Tobii Studio software were used to record the eye movements, utterances and mouse gestures of both participants concurrently. All participants were native Turkish speakers.

During the experiment, each pair was asked to collaboratively solve four different tangram puzzles by using the Tangram Simulator software (Spanger et al., 2009, 2010; Tokunaga et al., 2010). Figure 1 above displays a screen shot of the Tangram Simulator. Tangram puzzles require solvers to construct a target shape by using seven pieces, which include two large triangles, two small triangles, one medium-size triangle, a square and a parallelogram. Participants used mouse gestures to move and rotate the Tangram pieces to construct the desired shape on their shared workspace. Before the experiment, each participant was asked to complete a short training task to get familiarized with the puzzle interface.

Participants were assigned to either the role of the *operator* or the *instructor* during each task. The operator had the control of the mouse, but had no access to the goal shape. Only the instructor could see the target shape, so it was the instructor's job to guide the operators' actions by uttering instructions. After completing the first two tasks, participants switched their roles. The operator's mouse pointer was not visible to the instructor. In other words, the instructor could only see a change on the shared space if the operator actually moves or rotates a specific piece.

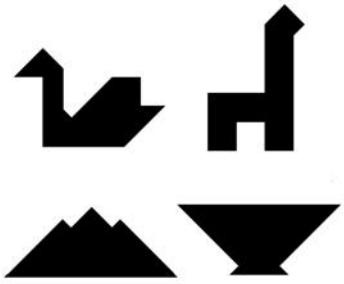

Figure 2: Swan, chair, mountain and vase constituted the four target shapes used during the experiment.

A total of four target shapes were used during the experiment (Figure 2). Pairs were allowed a maximum time of 15 minutes to work on each target. A hint was automatically provided by the software in every 5 minutes. The hint revealed the correct location of a single piece on the target description screen, so it was only visible to the instructor. The total duration of each experiment was approximately an hour.

In short, the experiment is particularly engineered in an effort to encourage participants to use referring expressions to coordinate their work. The roles assigned to the participants and the disembodied nature of the task were the two main constraints imposed by the activity. Hence, the task design eliminated the possibility of using cues such as pointing gestures and bodily orientations. The use of such interactional resources is beyond the scope of this corpus. A visual task that requires spatial arrangements of relevant objects was deliberately chosen to increase the chances of observing the use of referring expressions.

2.2 Annotation Process

The screen recordings of pairs, with overlaid eye-gaze data, were synchronized and transcribed with the Transana data analysis software (www.transana.org).

Figure 3 shows a snapshot of the Transana transcription interface. Two native Turkish annotators (one annotator was one of the authors) independently annotated eight dialogues from two pairs following the annotation guidelines provided by Spanger et al. (2010) with minor modifications with respect to the classification of the annotation tags (explained below). Accordingly, we focused on annotating noun phrases that referred to the pieces in the working space of the puzzle interface. An inter-annotator reliability analysis of the annotation scheme was conducted on a sample of 5,620 tokens extracted from the corpus. Two annotators independently annotated this sample by identifying which tokens constitute referring expressions, and then selecting an appropriate label from the annotation scheme. Holsti's (1969) method found that the percent agreement among the two annotators was 0.847, where 1 corresponds to perfect agreement. This method takes into consideration the number of disagreements among annotators in terms of which tokens should be annotated, but it fails to correct for the role of chance agreement.

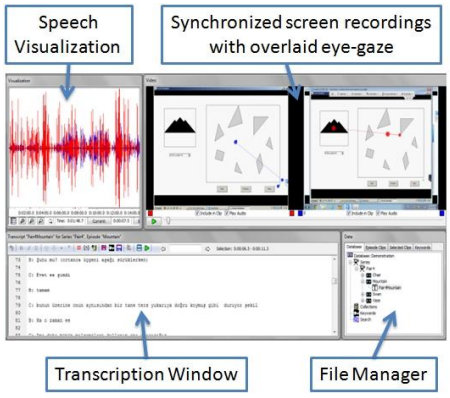

Figure 3: Interface of the Transana transcription tool, which displays a synchronized view of both participants' screens with overlaid eye gaze data.

3. Exploratory Analysis of the Corpus

In eight dialogues, we identified 1,109 tokens (844 produced by the solvers and 265 produced by the operators) with 170 different types of referring expressions. The collection of referring expressions involved 132 multiple-word referring expressions (418 tokens) and 38 single-word (691 tokens) referring expressions. Table 1 shows a partial list of referring expressions ordered by token frequency.

The data presented in Table 1 show that the participants produced the three demonstrative pronouns in Turkish (*o* 'it/that', *bu* 'this' and *şu* 'that') more frequently than others. As a consequence of the nature of the puzzle environment, they employed shape attributes (*parelelkenar* 'parallelogram', *üçgen* 'triangle' and *kare* 'square') in referring expressions. Finally, the participants used size attributes (*büyük* 'large', *küçük* 'small' and *orta boy* 'middle size'), demonstrative adjectives (*o* 'it/that', *şu* 'that' and *bu* 'this') and their combination (*o büyük üçgen* 'that large triangle') for modification of the shape attributes.

Table 1: Frequently used referring expressions in the corpus.

Referring Exp.	**%**	**Referring Exp.**	**%**
o 'it/that'	20.6	*küçük üçgen* 'small square'	3.9
bu 'this'	9.1	*orta boy üçgen* 'middle-size triangle'	2.9
şu 'that'	8.9	*o üçgen* 'that triangle'	2.6
paralelkenar 'parallelogram'	6.0	*şu üçgen* 'that triangle'	1.2
büyük üçgen 'big triangle'	5.7	*bu üçgen* 'this triangle'	1.1
üçgen 'triangle'	5.3	*şu paralelkenar* 'that parallelogran'	0.9
kare 'square'	5.0	*o büyük üçgen* 'that large triangle'	0.8

A more detailed analysis of syntactic/semantic properties of the referring expressions was conducted by a word-by-word based analysis of the identified referring expressions. For this, we annotated the single-word referring expressions and each word in the multi-word referring expressions according to their syntactic/semantic features. The feature list was prepared following the feature list identified by Spanger et al. (2009) for the Japanese corpus. We modified the feature list according to our findings peculiar to Turkish. The feature list is presented in Table 2.

Table 2: Syntactic/semantic features of the referring expressions

Feature	**Example**
Demonstrative	
Adjective	*bu üçgen* 'this triangle'
Pronoun	*bu* 'this', *şu* 'that', *o* 'it/that'
Nominalized form	*küçükler* 'small-PLU'
Partitive	*-den biri* 'one of …'
Determinative	*diğeri* 'other', *aynısı* 'same'
Pronominal Quantifier	*bu şey* 'this thing'
Attribute	
Size	*büyük üçgen* 'large triangle'
Shape	*büyük üçgen* 'large triangle'
Direction	*sola bakan* 'the one facing to left'
Spatial relation	
Projective	*sağdaki* 'the one on the right'
Topological	*dışarıdaki* 'the one outside'
Overlapping	*üstündeki* 'the one on the top'
Action mentioning	*çevirdiğin* 'the one you turned'
Time adverbial	*deminki* 'the one a moment ago'

The difference between Spanger et al.'s (2009) feature list and ours is the addition of a set of features (*nominalized form*, *partitive*, *pronominal quantifier* and *time adverbial*) in the Turkish feature list. The *determinative* class is involved in the 'other' category in the feature list for the Japanese corpus. We found it necessary to identify those features separately because we observed that the token frequency of those features in Turkish was higher than some of the common features in the Turkish corpus and in the Japanese corpus.

A frequency analysis of syntactic/semantic features of referring expressions revealed that the two types of attributes (shape and size) and the two types of demonstratives (adjectives and pronouns) were more frequently produced by the participants compared to other features. Those four types constitute approximately 84% of all the referring expressions produced by the participants.

The analysis revealed similarities between Turkish and Japanese referring expressions, as well. The major finding for the similarity between the two languages is that those four types of referring expressions were more frequent in the Japanese corpus, as well, constituting 85% of all the referring expression tokens. Table 3 gives a complete list of Turkish referring expressions, as well as Japanese referring expressions ordered by the percentage of token frequency.

Table 3: The syntactic/semantic feature distribution of the referring expressions (TR: Turkish, JAP: Japanese by Spangler et al., 2009)

Syntactic/Semantic Feature	**TR %**	**JAP %**
Shape (Attribute)	34.4	32.0
Pronoun (Demonstrative)	26.0	29.2
Size (Attribute)	14.4	14.1
Adjective (Demonstrative)	10.5	10.4
Determiner (Demonstrative)	4.69	1.59
Projection (Spatial Relation)	2.65	0.76
Action Mentioning	2.04	4.50
Partitive (Demonstrative)	1.56	NA
Pronominal Quantifier (Demonstrative)	1.02	NA
Nominalized Form (Demonstrative)	0.09	NA
Topological (Spatial Relation)	0.09	0.01
Direction (Attribute)	0.04	0.03
Time adverbial	0.04	NA
Overlap (Spatial Relation)	0.00	0.01

A comparison of the Japanese corpus with an English corpus of referring expressions was performed by Tokunaga et al. (2010), suggesting that Japanese participants use more *projection* (spatial relation) expressions and more *action mentioning* expressions compared to English participants. Our results suggest that Turkish participants exhibit a similar pattern with Japanese rather than English. However, a comparative investigation of the three languages will be performed after the completion of the analysis of all the recorded dialogues.

4. Conclusion and Future Work

In this paper, we presented the initial findings of an ongoing research on the construction of a corpus of Turkish referring expressions that employed a situated dialog environment. In its recent form, the data have been partially annotated.

Our preliminary results reveal that demonstrative pronouns, shape attributes and size attributes are the frequently employed features in referring expressions in the described situated dialogue environment of the Tangram puzzle solving task. The results also indicate that there are similarities between the syntactic/semantic feature distribution of Turkish and Japanese referring expressions. Like Japanese speakers, Turkish speakers also tend to use more projection and action mentioning referring expressions as compared to English speakers. We also identified additional features that are peculiar to Turkish referring expressions used in situated dialogs. Nevertheless, our findings are limited to the part of our corpus that has been annotated. A more thorough comparative investigation of the three languages will be performed once the annotation of all the recorded dialogues in our corpus is complete.

In the future we plan to expand this work across various dimensions. First, we will investigate whether the distribution of referring expression types differ across pairs, roles and task types (e.g. symmetric versus asymmetric target shapes). Second, we will focus on the eye tracking data to investigate how eye-gaze patterns are aligned with the referring expressions used by the participants. Finally, we will focus on the sequential organization of utterances that contain referring expressions to identify their communicational roles for the establishment and management of common ground for collaborative work. In particular, we aim to observe how different types of referring expressions are used in repair sequences to address problems of referential understanding.

Acknowledgements. The authors would like to thank Dr. Takenobu Tokunaga for making the Tangram Simulator software available to us, to Emine Eren and Semra Küçük for their help with corpus annotation, to Dr. Kürşat Çağıltay, Özge Alaçam and the METU HCI Research and Application Laboratory for granting us access to their eye tracking facilities, and to the anonymous reviewers for their comments and suggestions.

5. References

Byron, D., Mampilly, T., Sharma, V. & Xu, T. (2005). Utilizing visual attention for cross-modal coreference interpretation. In *Proceedings of CONTEXT 2005*, pp: 83-96.

Byron, D. & Fosler-Lussier, E. (2006). The OSU Quake 2004 corpus of two party situated problem solving dialogs. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC 2006)*.

Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22:1-39.

Di Eugenio, B., Jordan, P. W., Thomason, R. H. & Moore, J. D. (2000). The agreement process: An empirical investigation of human-human computer mediated collaborative dialogues. *International Journal of Human-Computer Studies*, 53(6): 1017-1076.

Goodwin, C. (1994). Professional Vision. *American Anthropologist* 96(3): 606-33.

Gundel, J. K. & Hedberg, N. (2008). *Reference: Interdisciplinary Perspectives.* New York, NY: Oxford University Press.

Hanks, W. (1992). The indexical ground of deictic reference. In A. Duranti & C. Goodwin (Eds.), Rethinking context: Language as an interactive phenomenon (pp. 43-76). New York: Cambridge University Press.

Hanks, W. (2009). Fieldwork on deixis. *Journal of Pragmatics*, 41, pp. 10-24.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Kılıçaslan, Y., Güner, E. S., & Yıldırım, S. (2009). Learning-Based Pronoun Resolution for Turkish with a Comparative Evaluation. *Computer Speech and Language*, 23(3): 311-331.

Say, B., Zeyrek, D., Oflazer, K., Özge, U. (2002). Development of a Corpus and a Treebank for Present-day Written Turkish. In *Proceedings of the Eleventh International Conference of Turkish Linguistics.*

Spanger, P., Yasuhara, M., Iida, R., Tokunaga, T. (2009). A Japanese Corpus of Referring Expressions Used in a Situated Collaboration Task. *In Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*. pp: 110-113.

Spanger, P., Yasuhara, M., Iida, R., Tokunaga, T., Terai, A. Kuriyama, N. (2010). REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*. (Dec). Online First, DOI: 10.1007/s10579-010-9134-8

Stoia, L., Shocley, D. M., Byron, D. K., & Fosler-Lussier, E. (2008). SCARE: A situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

Takenobu, T., Masaaki, Y., Asuka, T., Morris, D., & Belz, A. (2010). Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving. In *Proceedings of the Eighth Workshop on Asian Language Resources*. pp.38 – 46.

Tın, E. & Akman, V. (1994). Situated processing of pronominal anaphora. In *Proceedings of Second Conference for Natural Language Processing (KONVENS'94).* pp. 369–378.

Yüksel, Ö. & Bozşahin, C. (2002). Contextually Appropriate Reference Generation. *Natural Language Engineering*, 8: 69-89.

Zeyrek, D. Turan, Ü. D., Bozşahin, C., Çakıcı, R., Sevdik-Çallı, A., Demirşahin, I., Aktaş, B., Yalçınkaya, I., Ögel, H. (2009). Annotating Subordinators in the Turkish Discourse Bank. In *Proceedings of the ACL-IJCNLP, LAW Annotation Workshop III*. Singapore, August 6-7, 2009, pp: 44-48.

Building a Turkish ASR system with minimal resources

Arianna Bisazza and **Roberto Gretter**

Fondazione Bruno Kessler – Trento, Italy
bisazza@fbk.eu, gretter@fbk.eu

Abstract

We present an open-vocabulary Turkish news transcription system built with almost no language-specific resources. Our acoustic models are bootstrapped from those of a well trained source language (Italian), without using any Turkish transcribed data. For language modeling, we apply unsupervised word segmentation induced with a state-of-the-art technique (Creutz and Lagus, 2005) and we introduce a novel method to lexicalize suffixes and to recover their surface form in context without need of a morphological analyzer. Encouraging results obtained on a small test set are presented and discussed.

1. Introduction

Automatic Speech Recognition (ASR) systems are typically trained on manually transcribed speech recordings. Sometimes, however, this kind of corpora are either not available or too expensive for a given language, while it is pretty cheap to acquire untranscribed audio data, for instance from a TV channel. As regards language modeling (LM), only written text in the given language is required in principle. In reality, though, specific linguistic processings can be necessary to obtain reasonable performance in some languages. Turkish, with its agglutinative morphology and ubiquitous phonetic alternations, is generally classified as one of such languages. In this work, we investigate the possibility of building a Turkish ASR system with almost no language-specific resources. While this may seem an unrealistic scenario as more and more NLP tools and corpora are nowadays available for Turkish, we believe that our method may inspire further research on under-resourced languages with similar features, such as other Turkic languages or agglutinative languages in general.[1]

2. Unsupervised Acoustic Modeling

Acoustic modeling (AM) in state-of-the-art ASR systems is based on statistical engines capable to capture the basic sounds of a language, starting from an inventory of pairs ⟨utterance - transcription⟩. When only audio material is available, it can be processed in order to obtain some automatic transcription. Despite the fact that there will be transcription errors, it can be used to build a first set of sub-optimal AMs, which can in turn be used to obtain better transcriptions in an iterative way.

2.1. Audio recordings

International news are acquired from a satellite TV channel broadcasting news in different languages, including Turkish. It broadcasts a cyclic schema that lasts about 30 minutes, and roughly consists of: main news of the day (politics, current events); music & commercials; specialized services (stock, technology, history, nature); music & commercials. From an ASR perspective, data are not easy to handle, as several phenomena take place: often, in case of interviews, some seconds of speech in the original language are played before the translation starts; commercials are often in English; there is the presence of music; sometimes a particular piece of news may contain the original audio, in another language. In this paper we use 108 hours of untranscribed recordings (1 hour per day within almost 4 months) of the Turkish channel. Moreover, a small amount of disjoint audio data, about 12 minutes, was manually transcribed in order to obtain a test set (*TurTest*) containing 1494 reference words.

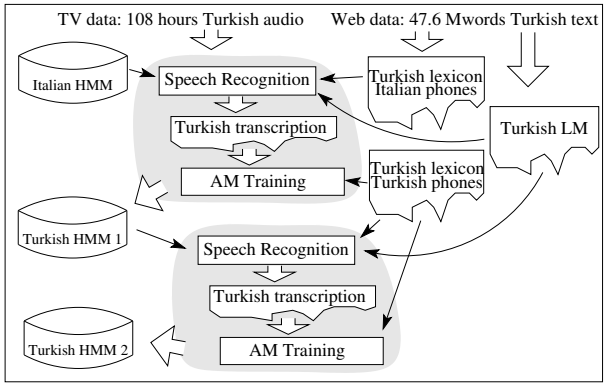

Figure 1: Block diagram of the procedure to bootstrap Turkish AMs from Italian ones.

2.2. Unsupervised acoustic training procedure

Figure 1 shows the unsupervised training procedure used for bootstrapping the phone Hidden Markov Models (HMMs) of a target language (Turkish) starting from those of a "well trained" source language (Italian) – for more details on this procedure see (Falavigna and Gretter, 2011). First we automatically transcribe the Turkish audio training data using a Turkish Language Model (LM), a lexicon expressed in terms of the Italian phones, and Italian HMMs. Then, a first set of Turkish HMMs (HMM 1 in Figure 1) is trained and used to re-transcribe the Turkish audio training data; this second transcription step makes use of a Turkish lexicon. A second set of Turkish HMMs (HMM 2 in Figure 1) is then trained using the new resulting transcriptions. Note that the procedure shown in Figure 1 could be iterated several times.

During the transcription stages, a Turkish LM was needed

[1]This work was partially funded by the European Union under FP7 grant agreement EU-BRIDGE, Project Number 287658.

REF:	ülkede**ki işçi sendikaları da** hükümetin **duyars**ız davrandığına **dikkati çekiyor**
HYP:	diğer i**ki işçi sendikaları da** internetten **duyar s**erdar arda **dikkati çekiyor**
REF:	**ülke çapında yapılan protesto gösterileriyle ma**denciler seslerini duyurmaya çalışırken
HYP:	**ülke çapında yapılan protesto gösterileri ile ma**vi jeans test edilmesi ve serkan

Table 1: Recognition of two Turkish utterances obtained with Italian acoustic models (first stage).

to drive the speech recognizer. It is coupled with a transcribed lexicon which provides the phonetic transcription of every word, expressed either in Italian phones (for the first iteration) or in Turkish phones (for the other iterations). Turkish phones which do not appear in the Italian inventory were mapped according to the following SAMPA table (http://www.phon.ucl.ac.uk/home/sampa/turkish.htm):

h:	h → ⟨sil⟩	*ɪ*:	1 → i	*ö*:	2 → o
ü:	y → u	*j*:	Z → dZ		

The collection of text data for training n-gram based LMs was carried out through web crawling. Since May 2009 we have downloaded, every day, text data from various sources, mainly newspapers in different languages including Turkish. A crucial task for LM training from web data is text cleaning and normalization: several processing steps are applied to each html page to extract the relevant information, as reported in (Girardi, 2007).

The LM for this stage was trained on 47.6 million words, which include the period of the audio recordings. Only number processing was applied at this stage. Perplexity (PP) on the small test set results very high (2508) while Out-of-Vocabulary (OOV) rate is reasonable (1.61%).

2.3. Convergence

Recognition on *TurTest* using the Italian AMs resulted in a 26.0% Word Accuracy (WA), corresponding to about 65% Phone Accuracy. Table 1 reports reference and ASR output for two samples, having 18 reference words and 14 ASR errors. Even if this corresponds to only 22.2% WA, phonetically more than half of the utterances are correct (highlighted in bold), resulting in a positive contribution to the AM training. The main causes of error at this stage were: acoustic mismatch, high perplexity and arbitrary phone mapping. However, despite the fact that 74.0% of the words are wrongly recognized, the second stage showed an encouraging 56.4% WA, which became 63.5% and 65.1% in the third and fourth stages.

3. Turkish Language Modeling

It is well known that morphologically rich languages present specific challenges to statistical language modeling. Agglutinative languages, in particular, are characterized by a very fast vocabulary growth. As shown for instance by Kurimo et al. (2006), the number of new words does not appear to level off even when very large amounts of training data are used. As a result, word segmentation appears as an important requirement for a Turkish ASR system. Two main approaches can be considered: rule-based and unsupervised. Rule-based segmentation is obtained from full morphological analysis, which for Turkish is typically produced by a two-level analyzer (Koskenniemi, 1984; Oflazer, 1994; Sak et al., 2008). On the other hand, unsupervised segmentation is generally learnt by algorithms based on the Minimum Description Length principle (Creutz and Lagus, 2005).

Another important feature of Turkish is rich suffix allomorphy caused by few but ubiquitous phonological processes. Vowel harmony is the most pervasive among these, causing the duplication or quadriplication of most suffixes' surface form. In this work we propose a novel, data-driven method to normalize (lexicalize) word endings and to subsequently predict their surface form in context. To our knowledge, this was only done by hand-written rules in past research.

3.1. Unsupervised Word Segmentation

Previous work (Arısoy et al., 2009) demonstrated that, for the purposes of ASR, unsupervised segmentation can be as good as, or even better than rule-based. Following these results, we adopt the unsupervised approach and, more specifically, the popular algorithm proposed by Creutz and Lagus (2005) and implemented in the Morfessor Categories-MAP software. The output of Morfessor for a given corpus is a unique segmentation of each word type into a sequence of morpheme-like units (*morphs*).

Instead of using each morph as a token, we follow a *'word ending'* (or *'half-word'*) approach, which was previously shown to improve recognition accuracy in Turkish (Erdoğan et al., 2005; Arısoy et al., 2009). In fact, while morphological segmentation clearly improves vocabulary coverage, it can result in too many small units that are hard to recognize at the acoustic level. As an intermediate solution between words and morphs, the sequence of non-initial morphs can be concatenated to form so-called *endings*. Note that the morphs do not necessarily correspond to linguistic morphemes and therefore a word ending can include a part of the actual stem.

Some examples are provided in Table 2. The segmentation of the first word (*saatlerinde*) is linguistically correct. On the contrary, in *çocukların*, the actual stem *çocuk* got truncated probably because the letter *k* is often recognized as a verbal suffix. The third word, *düşünüyorum*, is in reality composed of a verbal root (*düşün-*, 'to think') a tense/aspect suffix (*-üyor-*) and a person marker (*-um*). In this case, Morfessor included in the stem a part of the verbal tense suffix and oversplit the rest of the word. Finally, *diliyorum* was not segmented at all, despite being morphologically similar to the previous word. In any case we recall that detecting proper linguistic morphemes is not our goal and it is possible that statistically motivated segmentation be more suitable for the purpose of n-gram modeling.

The Morfessor Categories-MAP algorithm has an important parameter, the perplexity threshold (PPth), that regulates the level of segmentation: lower PPth values mean more aggressive segmentation. As pointed out by the software authors, the choice of this threshold depends on sev-

Word	Morfessor Annotation	Stem+Ending	Stem+Lex.Ending	Meaning
saatlerinde	saat/STM + ler/SUF + in/SUF + de/SUF	saat+ +lerinde	saat+ +lArHnDA	*in the hours of*
çocukların	çocu/STM + k/SUF + lar/SUF + ın/SUF	çocu+ +kların	çocu+ +KlArHn	*of the children*
düşünüyorum	düşünüyo/STM + r/SUF + u/SUF + m/SUF	düşünüyo+ +rum	düşünüyo+ +rHm	*I think*
diliyorum	diliyorum	diliyorum	diliyorum	*I wish*

Table 2: Chain of morphological processing on four training words. Morfessor annotation obtained with PPth=200.

eral factors, among which the size of the corpus. We then decided to experiment with various settings, namely PPth={100, 200, 300, 500}. Results will be given in Section 4. Morfessor was run on the whole training corpus dictionary, from which we only removed singleton entries.

3.2. Data-driven Morphophonemics

Vowel harmony and other phonological processes cause systematic variations in the surface form of Turkish suffixes, i.e. allomorphy[2]. For example, the possessive suffix *-(I)m* 'my' can have four different surface forms depending on the last vowel of the word it attaches to (ex.1-4), plus one if attached to a word that ends with vowel (ex.5):

1) *saç + (I)m* -> *saç**ı**m* 'my hair'
2) *el + (I)m* -> *el**i**m* 'my hand'
3) *kol + (I)m* -> *kol**u**m* 'my arm'
4) *göz + (I)m* -> *göz**ü**m* 'my eye'
5) *kafa + (I)m* -> *kafa**m*** 'my head'

As suffixes belong to close classes, we do not expect these phenomena to be the main cause of vocabulary growth. Nevertheless, we hypothesize that normalizing suffixes – or word endings in our case – may simplify the task of the LM and lead to more robusts models. Since the surface realization of a suffix depends only on its immediate context, we can leave its prediction to a post-processing phase.
In (Erdoğan et al., 2005) vowel harmony is enforced *inside* the LM by means of a weighted finite state machine built on manually written rules and exception word lists. More recently Arısoy et al. (2007) addressed the same problem by training the LM on lexicalized suffixes and then recovering the surface forms in the ASR output. This technique too required the use of a rule-based morphological analyzer and generator. On the contrary, we propose to handle suffix allomorphy in a data-driven manner. The idea is to define a few *letter equivalence classes* that cover a large part of the morphophonemic processes observed in the language. In our experiments we use the following classes:

$$A=\{a,e\} \quad H=\{ı,i,u,ü\}$$
$$D=\{d,t\} \quad K=\{k,ğ\} \quad C=\{c,ç\}$$

The first two classes address vowel harmony, while the others describe consonant changes frequently occurring between attaching morphemes. Note that defining the classes is the only manual linguistic effort needed by our technique. In the lexicalization phase, the letters of interest are deterministically mapped to their class, regardless of their context (see column 'Stem+Lex.Ending' in Table 2).
At the same time, a reverse index $\mathcal{I}$ is built to store surface forms that were mapped to a lexical form (very unlikely surface forms are discarded by threshold pruning). The LM is subsequently trained on text containing lexicalized endings and $\mathcal{I}$ is used to provide the possible pronunciation variants of each ending in the transcribed lexicon. After recognition, $\mathcal{I}$ is employed to generate the possible surface forms, which are then ranked by two statistical models assigning probabilities to ending surface forms in context. We assume that predicting the first 3 letters of an ending is enough to guess its complete surface form. As for conditioning variable, we use the full stem preceding the lexical ending if frequently observed, or else its last 3 letters only. This results in two models that are linearly combined: the *Stem Model* and the *Stem End Model*, respectively. The intuition behind this is that frequent exceptions to the generic phonological rules can be captured by looking at the whole stem, while for most of other cases knowing a small context is enough to determine an ending's surface form. Here is an example:

Stem Model		Stem End Model
p(+**lar**|kural)=.894	p(+lar|santral)=.026	p(+**lar**|*ral)=.242
p(+ler|kural)<.001	p(+**ler**|santral)=.308	p(+ler|*ral)=.200

Combination weights were set to $\langle 0.8,0.2 \rangle$ to give priority to the larger-context model (*Stem Model*). During post-processing, each lexical ending is assigned the surface form with the highest probability, among those provided by $\mathcal{I}$.

4. Experiments

Two text corpora were defined: *TurTrain* and *TurDev*. Both of them have been collected via web crawling, over two distinct periods (*TurTrain*: Jan 1, 2010 - Feb 15, 2012 and *TurDev*: Feb 16, 2012 - Feb 28, 2012). The same basic cleaning procedures were applied, in particular numbers were expanded (e.g. *2012 → iki bin on iki*) and punctuation was removed. *TurTrain* resulted in 129.9M words (lexicon size: 837K), while *TurDev* resulted in 3.2M words (99K).

4.1. Language Model Coverage and Perplexity

To evaluate the language modeling component of our ASR system, we measure OOV and PP on *TurDev* and on the reference transcription of *TurTest*, our ASR benchmark. In Table 3 the baseline word-level LM is compared with a series of LMs trained on *'word ending'* segmented data obtained with different PPth values. We recall that lower PPth means more aggressive segmentation by Morfessor. Note that perplexities are not directly comparable with one another, as the number of test tokens changes across settings.

4.2. Morphophonemic Normalization

With PPth equal to 200, the reverse index built on *TurTrain* contains 4355 ambiguous entries, i.e. lexicalized word endings with more than one surface form, and the average number of surface forms per entry is 2.3. To compute the accuracy of the surface form generator, we first lexicalize

[2]In this work we do not directly address stem allomorphy.

Preproc.	PPth	TurTrain #tokens	TurTrain lex.size	TurDev PP	TurDev OOV	TurTest PP	TurTest OOV
baseline		129.9M	837K	501	1.97	1442	0.94
morph	500	154.2M	733K	184	1.66	365	0.76
morph	300	161.4M	688K	148	1.59	260	0.72
morph	200	170.2M	636K	114	1.51	186	0.68
morph	100	173.5M	605K	105	1.48	169	0.66

Table 3: Impact of unsupervised *word ending* segmentation on number of training tokens and lexicon size; PP and OOV obtained on test sets by the corresponding 5-gram LMs.

the endings found in the development set, then we recover their surface forms in context by applying the models described in Section 3.2. Finally, we compare results with the original version of the text. We find that 27% of the tokens in *TurDev* are ambiguous lexicalized endings, and that 99.7% of them are assigned the correct surface form by our model. From a manual analysis, it also appears that some mismatches are actually due to the presence of wrong surface forms in the original text. In fact, misspellings are extremely common in web-crawled text (e.g. the non-Latin character *'ı'* replaced by *'i'*).

Given the very good performance reported, we integrate the model into our ASR system and measure its impact on language modeling. The OOV rate remains unchanged, but this is not surprising as lexicalization does not concern stems, which are the main responsible for vocabulary growth. Unfortunately, as shown in the row "lex" of Table 4, the effect on perplexity is also negligible.

	TurDev 4-gram	TurDev 5-gram	TurDev 6-gram
morph200	112.0	114.4	115.1
morph200,lex	112.1	114.0	114.6

Table 4: Effect of data-driven lexicalization on perplexity.

4.3. Speech Recognition

Speech recognition experiments were performed over *TurTest* using the same AMs described in Section 2. Table 5 reports results in terms of WA and, for the morphological case, Half-Word Accuracy (HWA). The latter simply corresponds to measuring WA *before* joining the half-words, which are the true output of the ASR system.

As a first observation, performance is reasonable and close to the state of the art, at least on our small test set. This is an important result, given that no language-specific resources were used on either the acoustic or language modeling side. Secondly, we compare the word-based approach (baseline) with the morphological approaches described above: WA improves from 71.55% to 73.69% (+2.14%) in the best experimental setting, that is 5-grams and PPth=200. In general we see that tuning the value of PPth is important as recognition accuracy varies significantly with it. Indeed, the intermediate values (300 and 200) yield the best performance overall. To our knowledge, previous work did not investigate this point but only used the default setting provided in the tool's distribution. Looking at HWA, trends are somehow different. However, it should be noted that here the number of reference units changes across settings, making values in different rows not directly comparable with one another. As regards the n-gram order, HWA figures confirm the trends observed on WA: 5-grams are better than both 4-grams and 6-grams.

From the last row of Table 5 we see that morphophonemic normalization has a negative effect on accuracy. This is in contrast with the improvements achieved by Arısoy et al. (2007) when applying a similar technique built on a rule-based morphological analyzer. Interestingly, though, the best result in the last row is obtained by the 6-grams, while in all other settings 5-grams are better. In future work we would like to investigate whether normalization can have a positive impact on 7-grams or even higher-order LMs.

	TurTest 4-gram	TurTest 5-gram	TurTest 6-gram
baseline	71.15| –	**71.55**| –	71.29| –
morph500	71.95|73.30	72.69|74.23	72.49|73.95
morph300	72.89|74.28	**73.69**|75.05	72.69|74.18
morph200	72.36|75.19	**73.69**|76.40	73.49|76.40
morph100	72.56|75.69	73.36|**76.87**	73.23|76.49
morph200,lex	71.69|74.42	72.09|74.86	**73.23**|**76.06**

Table 5: Recognition results in percentage word accuracy and half-word accuracy (WA|HWA).

So far we did not limit the vocabulary size. However, this is a parameter that tends to grow indefinitely with the size of the text corpus, and in our case reached 837K entries. Thus, we only keep the most frequent N entries, and test the effect on three parameters: WA, PP and OOV. Figure 2 reports the results, which highlight how the morphological approach is more robust to this effect as expected.

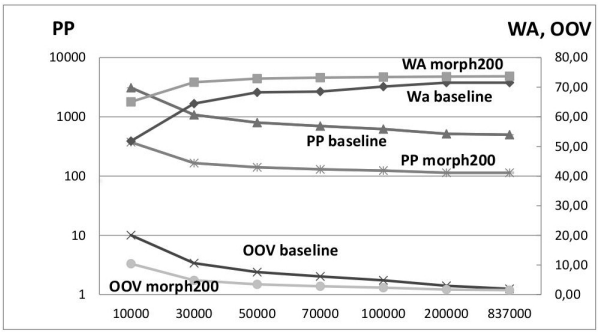

Figure 2: Results depending on lexicon size (*x*-axis).

5. Conclusions

We have shown how a Turkish ASR system with reasonable performance can be built without using language-specific resources: AMs were bootstrapped from those of a well-trained language, while unsupervised segmentation was applied to LM training data. The whole development cycle required only few minor interventions by an expert of the language. Experiments show that word-segmented models are more accurate and robust wrt lexicon size variations. Besides, WA appears to be notably affected by the degree of word segmentation. We have further presented a novel method to perform phonetic normalization of word endings. Intrinsic evaluation is very positive, however the effect on ASR is rather negative. While we plan to further investigate this effect, we hope that our work will inspire further research in under-resourced agglutinative languages.

6. References

Ebru Arısoy, Haşim Sak, and Murat Saraçlar. 2007. Language modeling for automatic turkish broadcast news transcription. In *Proceedings of INTERSPEECH*.

Ebru Arısoy, Doğan Can, Sıddıka Parlak, Haşim Sak, and Murat Saraçlar. 2009. Turkish broadcast news transcription and retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):874 –883.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*.

H. Erdoğan, O. Büyük, and K. Oflazer. 2005. Incorporating language constraints in sub-word based speech recognition. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 98–103.

Daniele Falavigna and Roberto Gretter. 2011. Cheap bootstrap of multi-lingual hidden markov models. In *Proceedings of INTERSPEECH*, pages 2325–2328.

Christian Girardi. 2007. Htmcleaner: Extracting relevant text from web. In *3rd Web as Corpus workshop (WAC3)*, pages 141–143.

Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *Proceedings of ACL*, pages 178–181.

Mikko Kurimo, Antti Puurula, Ebru Arısoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumäe, and Murat Saraçlar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 487–494.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *GoTAL 2008*, volume 5221 of *LNCS*, pages 417–427. Springer.

A prototype machine translation system for Tatar and Bashkir based on free/open-source components

Francis M. Tyers, Jonathan North Washington, Ilnar Salimzyanov, Rustam Batalov

Universitat d'Alacant, Indiana University, Kazan Federal University, Bashkir State University
Alacant, Bloomington, Kazan, Ufa
E-03071 Spain, IN 47405 (USA), Russia, Russia
ftyers@dlsi.ua.es, jonwashi@indiana.edu, ilnar.salimzyan@gmail.com, taqmaq@mail.ru

Abstract

This paper presents a prototype bidirectional machine translation system between Tatar and Bashkir, two minority Turkic languages of Russia. While the system has low open-domain coverage, results are presented that suggest that high accuracy may be obtained between these two closely-related languages, on a par with similar systems.

Keywords: Tatar, Bashkir, MT, Free software, Open-source

1. Introduction

This paper presents a prototype shallow-transfer rule-based machine translation system between Tatar and Bashkir, two closely-related minority Turkic languages of Russia.

The paper will be laid out as follows: Section 2. gives a brief description of the two languages; Section 3. gives a short review of some previous work in the area of Turkic–Turkic language translation; Section 4. describes the system and the tools used to construct it; Section 5. gives a very preliminary evaluation of the system; and finally Section 6. describes our aims for future work and some concluding remarks.

2. Languages

Tatar is a Turkic language spoken in and around Tatarstan by approximately 6 million people. Bashkir (Bashqort) is a Turkic language spoken by about 1.5 million people in and around Bashqortostan. Tatarstan and Bashqortostan are both republics within Russia. Both languages are co-official with Russian in their respective republics. The two languages belong to the same branch of the Kypchak group of Turkic languages. As they are very close relatives, they share many innovations, but Bashkir has quite a few phonological innovations beyond those of Tatar (such as rounding harmony and desonorisation of high-sonority suffix-initial consonants; cf. Washington (2010)) and the languages have a number of morphological differences (including different volitional participles). The spoken languages share a high level of mutual intelligibility, but many of the inherent similarities are obscured by their fairly different orthographical systems along with the phonological and morphological differences between the languages.

Aside from native speaker intuition, we also consulted the Bashkir Grammar of M. G. Usmanova (Усманова, 2006).

3. Previous work

Several previous works on making machine translation systems between Turkic languages exist, although to our knowledge none are publically available.

For systems between Turkish and other Turkic languages, there have been, for example, systems reported for Turkish-Crimean Tatar (Altintas, 2001b), Turkish-Azerbaijani (Hamzaoğlu, 1993), Turkish-Tatar (Gilmullin, 2008), and Turkish-Turkmen (Tantuğ et al., 2007).

4. System

The system is based on the Apertium machine translation platform (Forcada et al., 2011).[1] The platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other, more distantly related, language pairs. The whole platform, both programs and data, are licensed under the Free Software Foundation's General Public Licence[2] (GPL) and all the software and data for the 30 supported language pairs (and the other pairs being worked on) is available for download from the project website.

4.1. Architecture of the system

The Apertium translation engine consists of a Unix-style *pipeline* or *assembly line* with the following modules (see Fig. 1):

- A *deformatter* which encapsulates the format information in the input as *superblanks* that will then be seen as blanks between words by the other modules.
- A *morphological analyser* which segments the text in surface forms (SF) (*words*, or, where detected, multiword lexical units or MWLUs) and for each, delivers

[1] http://www.apertium.org
[2] http://www.fsf.org/licensing/licenses/gpl.html

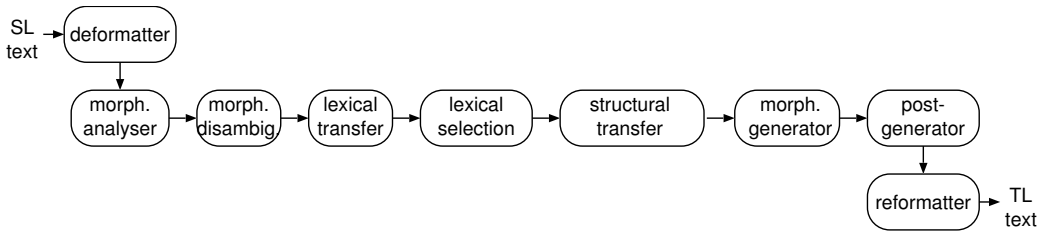

Figure 1: The pipeline architecture of the Apertium system.

one or more *lexical forms* (LF) consisting of *lemma*, *lexical category* and morphological information.

- A *morphological disambiguator* (constraint grammar) which chooses, using linguistic rules the most adequate sequence of morphological analyses for an ambiguous sentence.
- A *lexical transfer* module which reads each SL LF and delivers the corresponding target-language (TL) LF by looking it up in a bilingual dictionary encoded as an FST compiled from the corresponding XML file. The lexical transfer module may return more than one TL LF for a single SL LF.
- A *lexical selection* module which chooses, based on context rules the most adequate translation of ambiguous source language LFs.
- A *structural transfer* module which performs local syntactic operations, is compiled from XML files containing rules that associate an *action* to each defined LF *pattern*. Patterns are applied left-to-right, and the longest matching pattern is always selected.
- A *morphological generator* which delivers a TL SF for each TL LF, by suitably inflecting it.
- A *reformatter* which de-encapsulates any format information.

4.2. Morphological transducers

The morphological transducers are based on the Helsinki Finite State Toolkit (Linden et al., 2011), a free/open-source reimplementation of the Xerox finite-state toolchain, popular in the field of morphological analysis. It implements both the **lexc** formalism for defining lexicons, and the **twol** and **xfst** formalisms for modeling morphophonological rules. It also supports other finite state transducer formalisms such as **sfst**. This toolkit has been chosen as it – or the equivalent XFST – has been widely used for other Turkic languages (Çöltekin, 2010; Altintas, 2001a; Tantuğ et al., 2006), and is available under a free/open-source licence.

The morphologies of both languages are implemented in lexc, and the morphophonologies of both languages are implemented in twol.

Use of lexc allows for straightforward definition of different word classes and subclasses. For example, Tatar (but not Bashkir) has two classes of verbs: one which take a harmonised high vowel in the infinitive (the default), and one which take a harmonised low vowel in the infinitive. This was implemented in lexc with two similar continuation lexica for verbs: one pointing at a lexicon with an A-initial infinitive ending, and another pointing at a lexicon with an I-initial infinitive ending.

Use of twol allows for phonological processes present in the languages, like vowel harmony and desonorisation, to be implemented in a straightforward manner. For example, in Tatar, the A and I archiphonemes found in the infinitive are harmonised to one of two vowels each, depending on the value of the preceding vowel; the basic form of this process can be implemented in one twol rule.

The same morphological description is used for both analysis and generation. To avoid overgeneration, any alternative forms are marked with one of two marks, `LR` (only analyser) or `RL` (only generator). Instead of the usual compile/invert to compile the transducers, we compile twice, once the generator, without the `LR` paths, and then again the analyser without the `RL` paths.

4.3. Bilingual lexicon

The bilingual lexicon currently contains 2,834 stem to stem correspondences and was build by hand by a bilingual speaker of Tatar and Bashkir, translating a frequency list of the Russian National Corpus[3] into both languages in a spreadsheet. This spreadsheet was then converted into the Apertium XML dictionary format.

Entries consist largely of one-to-one stem-to-stem correspondences with part of speech, but also include some entries with ambiguous translations (see e.g., Fig. 2).

4.4. Disambiguation rules

The system has a morphological disambiguation module in the form of a Constraint Grammar (CG) (Karlsson et al., 1995). The version of the formalism used is vislcg3.[4]

The grammar currently has only four rules, but given the closeness of the languages, the majority of ambiguity may be passed through from one language to the other.

[3] http://ruscorpora.ru/en/

[4] http://beta.visl.sdu.dk/constraint_grammar.html

<e> <l>**юнәлеш**<s n="n"/></l><r>**йүнәлеш**<s n="n"/></r> </e>
<e> <l>**борын**<s n="n"/></l><r>**танау**<s n="n"/></r> </e>
<e> <l>**борын**<s n="n"/></l><r>**морон**<s n="n"/></r> </e>
<e> <l>**ераклык**<s n="n"/></l><r>**алыҫлыҡ**<s n="n"/></r> </e>
<e> <l>**ераклык**<s n="n"/></l><r>**йыраҡлыҡ**<s n="n"/></r> </e>

Figure 2: Example entries from the bilingual transfer lexicon. Tatar is on the left, and Bashkir on the right

4.5. Lexical selection rules

Likewise, lexical selection is not a large problem between Tatar and Bashkir, but a number of rules can be written for ambiguous words; for example, the Tatar word *борын* 'nose (person), nose (ship)' can be translated into Bashkir as either *танау* 'nose (person)' or *морон* 'nose (ship)'. A lexical selection rule chooses the translation *танау* if the immediate context includes a proper name.

Another example is the word *катлаулы* 'layered'. It is always translated to Bashkir as *ҡатмарлы*, except in the collocaton *катлаулы мәсьәлә* 'difficult matter/problem', which is translated as *ҡатлаулы мәсьәлә*.

5. Evaluation

Lexical coverage of the system is calculated over a freely available corpus of Bashkir, the Bashkir Wikipedia,[5] and over two freely available corpora of Tatar, the Tatar Wikipedia[6] and the New Testament in Tatar. The version of the translation tested was `r37137` from the Apertium SVN.[7] As shown in Table 2, the coverage is still far too low to be of use as a general broad-domain MT system, but we hope that it shows that a good proportion of the morphology of both languages is in place.

To get an idea of the kind of performance that could be expected from the system, we translated a simple story from Tatar to Bashkir and vice versa. The story may be found online,[8] and was used for pedagogical purposes in a recently workshop on MT for the languages of Russia.

Table 3 presents the Word Error Rate, an edit metric based on the Levenshtein distance (Levenshtein, 1966). This measure was calculated once all the stems in the text had been added to the system, thus presents an upper bound on the current performance of the transfer lexicon, and the disambiguation and transfer rules. The difference in the number of unknown words between translating Tatar→Bashkir and vice versa is because certain forms were not found due to lack of corresponding morphophonological rules.

We calculate the WER instead of other MT evaluation metrics such as BLEU as the WER is geared towards a particular task, that of measuring postedition effort. The translations of the story into Tatar and Bashkir were done in parallel to make them as close as possible, so using BLEU would give an over-optimistic view of the quality.

5.1. Error analysis

The majority of errors are currently due to mistakes and gaps in the morphophonology component; some minor problems still remain involving:

- Combinations of case and possessive suffixes,
- Orthographical representations of phonology,
- Vowel harmony processing on clitics (e.g., *да*/*дә* 'and') after unknown words.

6. Concluding remarks

To our knowledge we have presented the first ever MT system between Tatar and Bashkir, and the first ever MT system involving Bashkir. The system is available as free/open-source software under the GNU GPL and the whole system may be downloaded from SVN.[9]

We plan to continue development on the pair; the main work will be expanding the dictionaries with new lists of stems, and providing bilingual correspondences. The long-term plan is to integrate the data created with other open-source data for Turkic languages in order to make transfer systems between all the Turkic language pairs. Related work is currently ongoing with Chuvash–Turkish and Turkish–Kyrgyz.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments in improving the paper. This work has been partially funded by Spanish Ministerio de Ciencia e Innovación through project TIN2009-14009-C02-01.

References

Altintas, Kemal (2001a). A Morphological Analyser for Crimean Tatar. *Proceedings of Turkish Artificial Intelligence and Neural Network Conference*.

Altintas, Kemal (2001b). Turkish To Crimean Tatar Machine Translation System. Master's thesis, Bilkent University.

[5] http://ba.wikipedia.org/; `bawiki-20111210-pages-articles.xml.bz2`

[6] http://tt.wikipedia.org/; `ttwiki-20111215-pages-articles.xml.bz2`

[7] https://apertium.svn.sourceforge.net/svnroot/apertium

[8] https://apertium.svn.sourceforge.net/svnroot/apertium/branches/xupaixkar/rasskaz

[9] https://apertium.svn.sourceforge.net/svnroot/apertium/nursery/apertium-tt-ba

(Tatar) Input	Һава бүген бик әйбәт, җылы гына.
Mor. analysis	^Һава/һава<n><attr>/һава<n><nom>$ ^бүген/бүген<adv>$ ^бик/бик<adv>/бик<n><attr>/бик<n><nom>$ ^әйбәт/әйбәт<adj>/әйбәт<adj><subst><nom>$^,/,<cm>$ ^җылы/җылы<n><attr>/җылы<n><nom>/җылы<adj>/җылы<adj><subst><nom>$ ^гына/гына<postadv>$^./.<sent>$
Mor. disambiguation	^Һава<n><nom>$ ^бүген<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$ ^җылы<adj>$ ^гына<postadv>$^.<sent>$
Lex. transfer (+ selection)	^Һава<n><nom>/һауа<n><nom>$ ^бүген<adv>/бөгөн<adv>$ ^бик<adv>/бик<adv>$ ^әйбәт<adj>/әйбәт<adj>$^,<cm>/,<cm>$ ^җылы<adj>/йылы<adj>$ ^гына<postadv>/ғына<postadv>$^.<sent>/.<sent>$
Struct. transfer	^һауа<n><nom>$ ^бөгөн<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$ ^йылы<adj>$ ^ғына<postadv>$^.<sent>$
Mor. generation	һауа бөгөн бик әйбәт, йылы ғына.

Table 1: Translation process for the phrase *Һава бүген бик әйбәт, җылы гына* ‘The weather today is very nice, it is very warm’.

Corpus	Tokens	Coverage
Tatar New Test.	163,603	72.04%
Tatar Wikipedia	37,123	70.19%
Bashkir Wikipedia	12,267	65.99%

Table 2: Naïve vocabulary coverage over the three corpora.

Corpus	Direction	Tokens	Unknown	WER
story	tt→ba	311	9	8.97%
	ba→tt	312	1	7.72%

Table 3: Word error rate and over the small test corpus.

Forcada, Mikel L., Ginestí-Rosell, Mireia, Nordfalk, Jacob, O’Regan, Jim, Ortiz-Rojas, Sergio, Pérez-Ortiz, Juan Antonio, Sánchez-Martínez, Felipe, Ramírez-Sánchez, Gema, & Tyers, Francis M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2), pp. 127–144.

Gilmullin, R. A. (2008). The Tatar-Turkish Machine Translation Based On The Two-Level Morphological Analyzer. In *Interactive Systems and Technologies : The Problems of Human-Computer Interaction*. Ulyanovsk, pp. 179–186.

Hamzaoğlu, Ilker (1993). Machine translation from Turkish to other Turkic languages and an implementation for the Azeri language. Master’s thesis, Bogazici University.

Karlsson, F., Voutilainen, A., Heikkilä, J., & Anttila, A. (1995). *Constraint Grammar: A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics—Doklady 10, 707–710. Translated from Doklady Akademii Nauk SSSR*, pp. 845–848.

Linden, Krister, Silfverberg, Miikka, Axelson, Erik, Hardwick, Sam, & Pirinen, Tommi (2011). *HFST—Framework for Compiling and Applying Morphologies*, vol. Vol. 100 of *Communications in Computer and Information Science*, pp. 67–85. ISBN 978-3-642-23137-7.

Tantuğ, A. Cüneyd, Adali, Eşref, & Oflazer, Kemal (2007). A MT system from Turkmen to Turkish employing finite state and Statistical Methods. In *Proceedings of MT Summit XI, Copenhagen, Denmark*.

Tantuğ, A. Cüneyd, Adalı, Eşref, & Oflazer, Kemal (2006). Computer Analysis of Turkmen Language Morphology. pp. 186–193. Advances in natural language processing, proceedings (LNAI).

Washington, Jonathan North (2010). Sonority-based affix unfaithfulness in Turkic languages. Master’s thesis, University of Washington.

Çöltekin, Çağrı (2010). A freely available morphological analyzer for Turkish. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, pp. 820–827.

Усманова, М. Г. (2006). *Грамматика башкирского языка для изучающих язык как государственный*. Уфа.

Turkish Discourse Bank: Ongoing Developments

Işın Demirşahin*, Ayışığı Sevdik-Çallı*, Hale Ögel Balaban†, Ruket Çakıcı*, and Deniz Zeyrek*

*Middle East Technical University
Ankara, Turkey
†İstanbul Bilgi University
İstanbul, Turkey
demirsahin@ii.metu.edu.tr, ayisigi@ii.metu.edu.tr, hogel@bilgi.edu.tr, ruken@ceng.metu.edu,tr, dezeyrek@metu.edu.tr

Abstract

This paper describes the first release of the Turkish Discourse Bank (the TDB), the first large-scale, publicly available language resource with discourse-level annotations for Turkish. The TDB consists of a sub-corpus of the METU Turkish Corpus (MTC), which is annotated for discourse connectives; their arguments, i.e., the text spans they bring together; modifiers of the connectives, and supplementary spans that provide details for the arguments. In this paper, we describe the features of the MTC and the sub-corpus on which the TDB is built. We provide information about the annotations and other contents of the first release of the TDB. Finally, we describe the ongoing developments including annotating the sense and the class of the connectives, and the morphological features of the nominalized arguments of subordinating conjunctives.

Keywords: Turkish, discourse bank, discourse connectives

1. Introduction

Turkish Discourse Bank (the TDB) is the first large-scale publicly available language resource with discourse level annotations for Turkish. Following the style of Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008), the discourse connectives are regarded as discourse-level predicates lexically anchoring discourse relations. The text spans that are connected by means of the connectives are regarded as the arguments of these discourse-level predicates. The annotations in the corpus include the discourse connectives, the modifiers and the arguments of the connectives, and supplementary materials for the arguments. In (1), a sample annotation from the TDB is given. The connective is underlined; the first argument is in italics and the second argument in bold face.

(1) *İnsanlar tabiattan eşit doğarlar.* Dolayısıyla **özgür ve köle ayrılığı olmamalıdır.**
People are born equal by nature. As a result, **there should be no such distinction as the freeman and the slave.**

The annotations were carried out using the tool designed specifically for the TDB (Aktaş, et al., 2010). The annotations were performed by either three independent annotators, or by a pair of annotators and an independent individual annotator (Zeyrek et al., 2010; Demirsahin et al, ms).

2. Contents of the First Release

The TDB is freely available to researchers, and can be requested from www.tdb.ii.metu.edu.tr. The first release of the TDB includes the raw text files, annotation files, annotation guidelines, and a browser.

2.1. Text Files

The TDB is built on a ~400,000-word sub-corpus of METU Turkish Corpus (the MTC) (Say et al., 2002). the MTC is a 2 million-word resource of post-1990 written Turkish from multiple genres. A total of 159 files, 83 columns and 76 essays were excluded from the TDB, because these genres lack the conventional paragraph structure and make extensive use of boldface. These characteristics were not transferred to the MTC, which might have interfered with the reliable interpretation of the discourse relations and the specification of the extent of the arguments.
For the rest of the genres, the TDB preserves the genre distribution of the MTC, as shown in Table 1.

	the MTC		**the TDB**	
Genre	**#**	**%**	**#**	**%**
Novel	123	15.63	31	15.74
Story	114	14.49	28	14.21
Research/Survey	49	6.23	13	6.60
Article	38	4.83	9	4.57
Travel	19	2.41	5	2.54
Interview	7	0.89	2	1.02
Memoir	18	2.29	4	2.03
News	419	53.24	105	53.30
Total	787	100	197	100

Table 1: Distribution of the genres in the MTC and the TDB

2.2. Annotations

For each annotated text span, the text and the offsets for the beginning and the end of the span are kept in a standoff XML file. All tags except NOTE denote text spans. The annotation files include the content text and the

beginning and end offsets for text spans. A sample XML tree for the connective span of (1) is provided in (2).

(2)
```
<Conn>
   <Span>
      <Text>dolayisiyla</Text>
      <BeginOffset>15624</BeginOffset>
      <EndOffset>15635</EndOffset>
   </Span>
</Conn>
```

The following subsections provide details for tree nodes and the note attribute.

2.2.1. CONN (Connective)

The discourse connective is regarded as an immediate discourse-level predicate (Webber and Joshi, 1998; Webber, 2004) with two abstract object arguments, i.e., events, states, possibilities, situations, facts, propositions, and projective propositions such as questions, commands and desires (Asher, 1993). Connectives that link non-abstract objects and sentential adverbs that modify a single abstract object rather than linking two abstract objects, are not annotated. Table 2 shows five most frequent discourse connectives, compared to their total instances in the TDB.

	Discourse connectives		Other uses		Total instances	
Conn	**#**	**%**	**#**	**%**	**#**	**%**
ve 'and'	2112	28.2	5389	71.8	7501	100.0
için 'because'	1102	50.9	1063	49.1	2165	100.0
ama 'but'	1024	90.6	106	9.4	1130	100.0
sonra 'later'	713	56.7	544	43.3	1257	100.0
ancak 'however'	419	79.1	111	20.9	530	100.0

Table 2: Percent of discourse connectives and other uses

In the first release of the TDB, only explicit connectives are annotated. The discourse connectives are gleaned from coordinating conjunctions, subordinating conjunctions and discourse adverbials (Zeyrek & Webber, 2008). In addition to these, phrasal expressions are also annotated. These are subordinating conjunctions that take a deictic argument, which resolves to an abstract object. For instance, the postposition *rağmen* 'despite, although' can either take a nominalized subordinate clause or a deictic element such as *bu* 'this', resulting in the phrasal expression *buna rağmen* 'despite this'. Although syntactically the argument of the postposition is the deictic element, the TDB annotations select the whole phrasal expression as the connective, and annotate the abstract object the anaphora resolves to as the argument, in order to more explicitly reflect the discourse relations between the abstract objects.

A total of 8483 relations are annotated in the TDB. The annotators searched for 77 tokens. This number includes various forms of one root, such as *amaçla* 'goal+INS' and *amacıyla* 'goal+POS+INS'. 143 distinct text spans were annotated as discourse connectives, including phrasal expressions and constructions based on a token. For instance, *buna rağmen* 'despite this', *bunlara rağmen* 'despite these', *herşeye rağmen* 'despite everything', are annotated as distinct connectives. Likewise, the token *yandan* 'side+ABL' returns *bir yandan* 'on one hand' and a a variety of phrases as its second part, such as *bir yandan da, diğer yandan, öbür yandan,* and *öte yandan*, all of which come to mean 'on the other hand'. Most variations of connectives can be collapsed to few common roots as exemplified in Table 3.

Root	**Variations**
amaç- 'goal'	bu amaçla, amacıyla, amacı ile
dolayı- 'because'	dolayı, dolayısıyla, dolayısı ile, bundan dolayı, bu sebepten dolayı
neden- 'reason'	bu nedenle, o nedenle, bu nedenlerle, yukarıdaki nedenlerle, nedeniyle, nedeni ile
sonuç- 'result'	sonuçta, sonucunda, sonuç olarak, bunun sonucunda, bunların sonucunda
zaman- 'time'	zaman, bir zamanda, aynı zamanda, o zaman, ne zaman…o zaman

Table 3: Some of the common roots for morphological varieties of connectives

2.2.2. MOD (Modifier)

The modifiers are spans that specify or intensify the meaning of the connective, or signify the modality of the relationship. For example, the discourse adverbial *sonra* 'later' can be modified for duration by *iki gün* 'two days' or the relation indicated by the subordinator *için* 'because/ for' can be modified for modality by *belki* 'perhaps'.

2.2.3. ARG1, ARG2 (First and Second Argument)

Similar to the PDTB, the argument that syntactically hosts the connective is called the second argument (ARG2) and the other argument is called the first argument (ARG1). Arguments of the discourse connectives can be single or multiple verb phrases, clauses or sentences, i. e., any text span with an abstract object interpretation.

2.2.4. SHARED (Shared Material)

The SHARED span was introduced to the TDB for the spans that belong to both Arg1 and Arg2 of a connective. A shared material may be the common subject, object or adjunct.

2.2.5. SUPP (Supplementary Material)

Supplementary materials are selected for the arguments or shared spans: SUPP1 for ARG1, SUPP2 for ARG2 and SUPP_SHARED for SHARED. These tags specify the spans of text necessary to fully interpret the arguments. In the TDB, the supplementary materials are extensively used to include the resolutions of discourse-level anaphora in the arguments.

2.2.6. NOTE

NOTE is an attribute of the relation tag, as in (3)[1].

(3) `<Relation note="" sense="" type="EXPLICIT">`

The annotators can enter free text in the notes field. This field is used for entering the rationale of the annotation, the problems annotators encountered during the annotation, or alternative annotations to the current one.

2.3. Annotation Guidelines

The annotation guidelines provide the definitions of key terms and general criteria for the annotations. The guidelines are supported with rich examples of both the annotated and unannotated cases.

2.4. The Browser

A browser specifically created for the TDB (Şirin, et al., 2012) is included in the first release. The browser enables the users to view all annotations on each file. The quick search feature enables the user to filter the files for connectives and genre. The advanced search feature offers the means to perform text and regular expression searches. A user manual is included in the distribution of the first release.

3. Ongoing Developments

Most discourse connectives have multiple uses. In the TDB, we have encountered connectives that can belong to multiple syntactic classes, such as subordinator and discourse adverbial. Also, most discourse connectives are polysemous to various degrees. In order to disambiguate such ambiguities, we introduce connective class and Arg2 feature annotations, as well as a PDTB-style sense annotation (Miltsakaki et al., 2005; Prasad et al, 2008).

3.1. CLASS (Connective Class)

The roots like *amaç-* 'goal', *neden-* 'reason', *netice-* 'result', *saye-* 'thanks to', and *yüz-* 'due to' may form subordinators and phrasal expressions. The subordinators are in the form root+POS+INS whereas their corresponding phrasal expressions have the form root+INS. However, the syntactic class of all such connectives cannot be figured out directly from the morphology of the connective. Some roots such as *sonuç-* 'result', form the subordinator *sonucunda* 'result+POS+LOC', as well as phrasal expressions, e.g. *bunun sonucunda* 'as a result of this'. Since phrasal expressions are annotated with the anaphoric expression in the text span, the connective class of *sonucunda* can be disambiguated from the CONN span. Still, there are connectives that are completely ambiguous in terms of subordinator and discourse adverbial uses, such as *sonra* in (5) and (4), respectively.

(4) **Sana aşık olduktan** sonra *karısından boşandı.*
He divorced his wife after **falling in love with you.**

(5) *Adam öldüğünü sandı,* **öldürüldüğünü** sonra.
The man thought he was dead; then (he thought) **that he was murdered.**

CLASS is a relation attribute like sense and notes. It has a limited set of values: CON for coordinating conjunctions, SUB for subordinating conjunctions, ADV for discourse adverbials and PHR for phrasal expressions. In addition, parallel constructions are marked with PAR together with the connective class of the compulsory item in the construction. For example, PAR CON for the parallel construction of the coordinating conjunctive *ya…ya* 'either…or', or PAR PHR for *ne zaman…o zaman* "when…then".

The preliminary connective class annotations have provided the connective class breakdown for the following ambiguous spans, given in Table 4[2]:

Span	Subordinating Conjunctive	Discourse Adverbial	Total
ardından 'following'	32	37	69
dolayısıyla 'as a result of'	2	64	66
önce 'first, before'	76	45	121
sonra 'than, later'	273	376	649

Table 4: Connective class disambiguation for ambiguous spans

3.2. ARG2FEAT (Feature Annotation for Second Arguments of Subordinators)

Most of the subordinating conjunctives in Turkish take nominalized clauses as their second arguments. These nominalizations can have a variety of morphological features, which makes the TDB a valuable source for studying nominalized abstract objects.

The morphological properties of the nominalized arguments also allow a further degree of disambiguation in case of *için* 'because, for'. *İçin* can express goal or cause driven relations. The sense of the relation can be disambiguated between goal and cause by simply looking at the morphology of the second argument. In (6), the *–mek için* marks a goal driven relation by taking an infinitival clause as argument, and in (7) *-dığım için* marks a cause driven relation by taking a factive clause (see also Table 5 below).

(6) **Onu görmek** için *tüm zamanınızı o parkta geçirmeye başlarsınız.*
In order to **see her** *you start to spend all your time in that park.*

[1] The first release of the TDB does not include sense annotation. The sense attribute of the relation tag is included to easily implement sense annotation in future releases and to ensure the compatibility of the sense tag with the current release of the browser.

[2] This table does not include parallel constructions and phrasal expressions including these spans, because their CONN spans already disambiguate their connective class; for instance the span *bunun ardından* 'following this' is unambiguously a phrasal expression as *ilk olarak...ardından* 'first…then' is a parallel construction.

(7) **Üvey babamı görmek istemediğim** için *yıllardır o eve gitmiyorum.*
Since **I don't want to see my step father,** *I haven't been to that house for years.*

Like CLASS, the ARG2FEAT is a relation attribute, which will be left blank for classes other than subordinating connectives.
A preliminary morphological annotation for (6) is INF which stands for infinitive, and for (7) FAC + AGR which stands for factive clause with person agreement. Other examples would be NOM MA + POS AGR + ABL CASE (nominalized with –mA, with person agreement on possessive case, attributed ablative case by the postposition) for … *olmalarından dolayı* 'although they are …', and CNV CA + DAT CASE (converb –cA, attributed dative case by the postposition) for *duyuncaya kadar* 'until hearing'.
Table 5 shows the disambiguation of *için* annotations in the TDB with respect to goal and cause driven relations.

Goal driven	
inf (-mAk) için	510
-mA + pos agr için	239
-mA için	6
-Iş + pos agr için	6
-Iş için	2
-Im + pos agr için	7
Goal Total	770
Cause driven	
-dIğI + agr için	276
- (A)cAğI + agr için	12
Cause total	288
İçin total	1058

Table 5: Goal - cause disambiguation for subordinator *için*

3.3. SENSE

Some connectives such as the subordinator *gibi* 'like, as/just as" cannot be disambiguated by morphology. *–dIğI gibi* marks an expansive relation in (8), a similarity relation in (9), and a temporal immediate succession relation in (10), with no morphological distinction on its argument.

(8) **Kahve değirmeninin nerede olduğunu bilmediği** gibi*, bulacağını da sanmıyordu*
In addition to **not knowing where the coffee mill is**, *he didn't think that he would be able to find it*.

(9) **Sizin yaptığınız** gibi *açık konuşacağım.*
I will speak frankly just like **you do.**

(10) **Bisikletine atladığı** gibi *pedallara basıyor.*
As soon as **he jumps on the bicycle**, *he hits the pedals.*

In addition to connectives like *gibi* that mark distinct sense classes such as EXPANSION and TEMPORAL relations, most connectives signal several types and subtypes of senses. For example, *ama* 'but' can signal CONTRAST, CONCESSION, EXCEPTION as well as PRAGMATIC variants of these senses.
For sense annotation we have taken the PDTB sense hierarchy (Prasad, 2007) as a starting point. Similar to Tonelli (2012), who discovered that the PDTB sense tags need to be expanded for spoken corpus annotations because of the extensive pragmatic uses, we have discovered that the rich variation of genres in the TDB calls for expansion of the sense hierarchy. In preliminary sense annotations, we have encountered a wide variety of pragmatic uses of *ama* 'but' including OBJECTION (11) and CORRECTION (12).

(11) - *Sana kahve yapacağım.* - Ama **çok içmedim.**
- *I will make you some coffee* - But **I haven't drunk much.**

(12) *Öyle bir kadın var!* Ama **o başkası!**
There is such a woman! But **she is someone else!**

The sense annotations are at a very early stage and the sense hierarchy is likely to be modified more as annotations progress.

4. Conclusion

In this paper we have introduced the features of the first release of the TDB. We also presented the ongoing developments for further enrichments, namely connective class annotation, Arg2 feature annotation and sense annotation.
The first two of these developments are well underway, and have already revealed detailed descriptives, such as the total connective class breakdown of disambiguated connectives in the TDB (Table 6). The number of distinct connectives increased from 143 to 150 (cf. § 2.2.1); because after the disambiguation processes, spans such as *ardından*-sub and *ardından*-adv or *için*-goal and *için*-cause are counted as distinct connectives.

		Single	**Parallel**	**Total**
Coordinating Conjunctive	Spans	15	12	27
	Relations	4348	129	4477
Subordinating Conjunctive	Spans	31	1	32
	Relations	2285	2	2287
Discourse Adverbial	Spans	32	18	50
	Relations	1152	73	1225
Phrasal Expression	Spans	40	1	41
	Relations	490	4	494
Total	Spans	118	32	150
	Relations	8275	208	8483

Table 6: Connective class breakdown of disambiguated connectives in the TDB

We believe that connective class, Arg2 feature, and sense annotations will contribute to the further study of Turkish in particular and provide a unique perspective to the studies in discourse in general.

5. Acknowledgements

We gratefully acknowledge the support of Turkish Scientific and Technological Research Council of Turkey (TUBITAK) and METU Scientific Research Fund (no. BAP-07-04-2011-005).

6. References

Asher, N. (1993). *Reference to Abstract Objects in Discourse.* Kluwer Academic Publishers.

Aktaş, B., Bozşahin, C., and Zeyrek, D. (2010). Discourse Relation Configurations in Turkish and an Annotation Environment. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV).*

Demirşahin, I., Yalçınkaya, İ., and Zeyrek, D. (ms). Pair Annotation: Adaption of Pair Programming to Corpus Annotation.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., and Webber, B. (2005). Experiments on sense annotation and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report 203, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, Pennsylvania.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation. (LREC'08).*

Say, B., Zeyrek, D., Oflazer, K., and Özge, U. (2002). Development of a Corpus and a Treebank for Present-day Written Turkish. In *Proceedings of the Eleventh International Conference on Turkish Linguistics (ICTL 2002).*

Şirin, U., Çakıcı, R., and Zeyrek, D. (2012). METU Turkish Discourse Bank Browser. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi. A. (2010). Annotation of Discourse Relations for Conversational Spoken Dialogues. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*.

Webber, B. (2004). D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, *28*(5), 751-779

Webber, B., and Joshi, A. (1998). Anchoring a lexicalized tree-adjoing grammar for discourse. In Stede, M., Wanner, L., Hovy, E. (Eds) *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pp. 86–92. Association for Computational Linguistics.

Zeyrek, D., and Webber, B. (2008). A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Turkish Corpus. In *Proceedings of the 6th Workshop on Asian Language Resources, The Third International Joint Conference on Natural Language Processing (IJNLP).*

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Balaban, H. Ö., Yalçınkaya, İ., and Turan. Ü. D. (2010). The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop* (*LAW IV).*

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.

The Outline of an Ottoman-to-Turkish Automatic Machine Transliteration System

Esma F. bilgin, e.f.bilgin@gmail.com
Atakan Kurt, akurt@fatih.edu.tr (Corresponding author)
Computer Eng. Dept., Fatih University, Istanbul, 34500, Turkey

Abstract

The Ottoman script is a writing system of the Turkish language which was in use from the early the 13th century until the 20th century. The transliteration of Ottoman script to Latin-based modern Turkish script is necessary in order to make a huge collection of documents available to readers. The transliteration problem can be reduced to pronunciation generation in Turkish for the Ottoman script, because the pronunciation of words remains the same. The main problem of the transliteration is the lack of a regular of orthography in the Ottoman script. The complexity of the problem requires a combination of NLP techniques beyond simple character mappings. This paper outlines the Ottoman orthography in general and discusses the complexities, problems, difficulties, exceptional cases in the Ottoman orthography. Then the vowel and consonant mappings between the two scripts are defined. Finally we present the outline of an automatic machine transliteration framework from Ottoman to Turkish currently under development.

1. Introduction

The Ottoman script is the writing system used by Turks from the early 13th century until the first half of the 20th century. The alphabet used for the Ottoman script is an extended version of 28-letter Arabic alphabet. Ottoman is written from right to left in a cursive manner as in the case of Arabic. *Ottoman Turkish* of the past, which is simply referred to as Ottoman henceforth, is an amalgamation of the Turkish language with words, phrases and some morphological/grammatical components borrowed from Arabic and Persian but its main components were still Turkish [Davids 1832, Barker 1854, Wells 1880, Redhouse 1884]. Many of the loanwords and structures became so localized that they eventually became an inseparable part of the *modern Turkish* in use today.

Although they are two different writing systems of the same language, the transliteration from the Ottoman script to the Turkish script is nontrivial. They have different alphabets and different orthographic rules. Turkish orthography is well defined and always obeyed in writing. However Ottoman orthography is complex and not well understood by many. It was not standardized over the centuries it was in use. What's more, there are too many exceptions, irregularities, and cliché forms to orthographic rules. [Kurt 1996, Develi 2006, Timurtaş 2003]

The pronunciation of a written word plays an intermediary role in transliteration. The pronunciation generation for a word is a complex and highly context dependent process. The Ottoman script always represents consonants while it usually lacks vowels. But the vowel omission in the Ottoman script does not seem to follow a regular orthography. Reader is expected to deduce missing vowels from the context. On the other hand the Turkish alphabet represents each phoneme with a single letter and therefore Turkish script has a straightforward spelling system.

Paper organization is as follows; Section 2 is dedicated to Ottoman and Turkish scripts and character mappings between them. Difficulties, problems and exceptional cases are summarized in Section 3. Section 4 introduces the framework of the automatic machine transliteration system. The approach is demonstrated with an example in Section 5. Section 6 is reserved for conclusions.

2 Ottoman and Turkish Scripts

Following table shows the character mappings between the Ottoman and Turkish letters. Note that there are 1-to-many mappings for some Turkish characters. In order to differentiate each element of such mappings the transcription characters are given as well.

Ottoman	Name	IPA	Transcription	Turkish	Ottoman	Name	IPA	Transcription	Turkish
ا	elif	/a/, /e/, /æ/	a, â, e	a, e	ض	dad	/d/, /z/	ż, ḍ	d, z
ء	hemze	/ʔ/	’	'	ط	tı	/t/	ṭ	t
ب	be	/b/, /p/	b, p	b	ظ	zı	/z/	ẓ	z
پ	pe	/p/	p	p	ع	ayın	/ʕ/	‘	'
ت	te	/t/	t	t	غ	gayın	/g/, /ɯ/	ġ	g, ğ

ث	se	/s/	s̲	s	ف	fe	/f/	f	f
ج	cim	/dʒ/, /tʃ/	c, ç	c, ç	ق	kaf	/k/	k̲	k
چ	çim	/tʃ/	Ç	ç	ك	kef	/c/, /ɟ/, /ɯ/, /ŋ/	k, g, ñ	k, g, ğ, n
ح	ha	/h/	ḥ	h	گ	gef (Kaf-i Farsi)	/ɟ/	g	g
خ	hı	/h/	h̲	h	ڭ	nef (Kaf-i Türki)	/ŋ/	ñ	n
د	dal	/d/	d	d	ل	lam	/ɬ/, /l/	l	l
ذ	zel	/z/	z̲	z	م	mim	/m/	m	m
ر	re	/ɾ/	r	r	ن	nun	/n/	n	n
ز	ze	/z/	z	z	و	vav	/v/, /o/, /ø/, /u/, /y/	v, o, ô, ö, u, û, ü	v, o, ö, u, ü
ژ	je	/ʒ/	j	j	ه	he	/h/, /e/, /æ/, /a/	h, e, a	h, e, a
س	sin	/s/	s	s	لا	lamelif	/ɬ a /, /l a /	lâ	la
ش	şın	/ʃ/	ş	ş	ى	ye	/j/, /ɯ/,/i/	y, ı, i, î	y, ı, i
ص	sad	/s/	ṣ	s					

Table 1 Character Mappings

The character correspondence table above and some exceptions not mentioned here due to space limitation shows that there are some 1-to-1, 1-to-many and many-to-many mappings between the Ottoman and Turkish characters making transliteration quite difficult. Mappings between vowels are conditional and not straightforward. On the other hand consonant mappings are relatively simpler but still problematic. 1-to-many, many-to-1 consonant relations are summarized in following tables.

Turkish	Ottoman
Ç	ج چ
D	د ض ط
G	غ گ
Ğ	غ ك
h	ه ح خ
k	ك ق
n	ن ڭ
p	ب پ
s	س ص ث
t	ت ط
z	ز ظ ض

Table 2 Character mappings – 1

Ottoman	Turkish
ب	b p
ج	c ç
ض	d z
ط	t d
غ	g ğ
ك	k ğ (g n)

Table 3 Character Mappings - 2

3. Problems arising from Ottoman Script in Transliteration

The evolution of Turkish language in centuries, the modifications on loanwords and the composition of words and structures of different languages introduced inconsistencies in the Ottoman spelling system. The source of irregularities and exceptional cases in the Ottoman orthography are the followings: Turkish origin words, the loanwords, the hybrid words and the noun adjuncts. Each word class introduces different kinds of problems which should be handled though different orthographic solutions. Due to space limitation this section is left out.

4. Ottoman-to-Turkish Machine Transliteration Framework

It is clear from the challenges mentioned above that transliteration between the Ottoman script and the Turkish script is multidimensional and beyond matching graphemes. Machine transliteration is drawing more attention recently. [Linden 2006, Halpern 2007, Malih 2008, Saini 2008, Jawaid et al 2009, Karimi et al 2011] Previous attempts at transcribing and/or transliteration from Ottoman had only limited success on small specific input texts [Emci 1990, Şişman 1995]. This paper suggests a solution which brings different NLP

techniques together. Steps in this approach includes morphological parsing/generation, bilingual lexicons for stems and suffix clusters, spellchecking for misspelled words, word boundary detection for misplaced space/ZWNJ characters, using n-gram statistics for exploiting contextual information in word disambiguation and intralingual translation to some extent. Following solution scheme lists these steps:

1. Morphologic Parsing is the process of extracting possible stem and suffix pairs in a given Ottoman word.
2. Transliteration Dictionary is used for looking-up extracted stem and suffix pairs in bilingual stem and suffix dictionaries respectively to retrieve the matching Turkish stem and suffix pairs.
3. Morphological Synthesis is the reconstruction of possible Turkish word transcriptions, and in turn, transliterations out of stem and suffix pairs.
4. Word Disambiguation is a method for choosing the right word via ranking the produced Turkish words with n-gram statistics.
5. Unrecognized Words (Error Handling) is the handling typographical errors, word segmentation errors and unknown words by spell checking/correction and vowelization.
6. Detecting noun adjuncts is done in post-processing phase.

Bilingual transliteration dictionaries for stems and suffix clusters are employed in the framework as examplified in tables below.

Origin	Turkish	Ottoman
Ara.	facir	فاجر
Tur.	faça	فاچه
Tur.	façuna	فاچونه
Per.	fağfur	فغفور
Per.	fağfurî	فغفوري

Table 4 Stem Dictionary

Turkish	Person	Ottoman
HndAkH	P2	إكدەكى
HndAkH	P3	ؤندەكى
HndAkHnA	P2	كدەكىنە
HndAkHnA	P3	ؤندەكىنە
HndAkHndA	P2	كدەكىندە

Table 5 Suffix Clusters

5. Demonstration

This subsection gives examples for each step of the proposed solution. It should be noted that the steps may not apply in order. The order of subtasks depends on the actual case at hand. Consider the following string for the demonstration of the algorithm:

بيكْ آتلى آقينلرده چو جوكلر… bin atlı akınlarda ço cuklar…

Below we give a word by word transliteration of this phrase to Turkish this approach. Note that there is typo, a space, in the last word.

Input	Steps	Output
بيكْ	1- Morphological parsing 2- Look up	بيكْ بيكْ => bin
آتلى	1- Morphological parsing 2- Look up 3- Morphological synthesis	آتلى => آت + لى آت => at لى => lH at + lH => atlı
آقينلرده	1- Morphological parsing 2- Look up 3- Morphological Synthesis * invalid suffix 2- Look up 3- Morphological Synthesis	1) آقينلرده => آق + ينلر + ده 2) آقينلرده => آقين + لر + ده 1) آق => ak نلر=> ينلرÂ ي => yAnlAr ده => DA yAnlAr => yanlar => Error* DA => da 2) آقين => akın لر => lAr ده => DA akın + lAr + DA => akınlarda
چو	1- Morphological parsing 5- Error 1- join with the previous 2- join with the next	Failure 1) آقينلردەچو => no stem 2)چوجوكلر => لر + چوجوك

	3- Morphological Synthesis	çocuk + lAr => çocuklar => چوجوكلر

6. Conclusions

The system introduced above has a dictionary-based approach to the transliteration problem. The irregularities in the pronunciation of the Ottoman script justify this approach. As shown above, the problem is not trivial and the solution is not simple. A number of NLP techniques including morphological parsing and generation, word segmentation, spell checking/correction, n-grams are being used to build an automatic machine transliteration system. As the system relies on bilingual lexicon, any improvement in the transliteration dictionary would increase both performance and accuracy. Current dictionary includes around 30.000 words which should be tolerably sufficient for general texts like newspapers and magazines.

References

Barker, W.B., 1854, A Practical Grammar of the Turkish Language: With Dialogues and Vocabulary, publisher B. Quaritch.

Develi, H., 2006, *A Guide for Ottoman Turkish Volume 1 [in Turkish]*, 8th edition, Kitabevi, İstanbul.

Emci, N., 1990, *Automatic Ottoman Text Transcription*, M.S. Thesis, Middle East Technical University.

Halpern, J., 2007, "The Challenges and Pitfalls of Arabic Romanization and Arabization", *The Second Workshop on Computational Approaches to Arabic Script-based Languages (CAASL2)*, Stanford University.

Jawaid, B. and T. Ahmed. *2009,* "Hindi to Urdu conversion: beyond simple transliteration", *Proceedings of Conference on Language and Technolog*, Lahore.

Karimi, S., F. Scholer and A. Turpin, 2011, "Machine Transliteration Survey", *ACM Computing Surveys*, Vol 43, Issue 3.

Kurt, Y., 1996, *Ottoman Lessons [in Turkish]*, 3rd edition, Akçağ Yayınları, Ankara.

Lindén, K., 2006, "Multilingual Modeling of Cross-Lingual Sğelling Variants", *Information Retrieval*, Vol. 9, No. 3.

Davids, A. L., 1832, A grammar of the Turkish language, published by Parbury & Allen, London.

Malik, M. G. A., C. Boitet and P. Bhattacharyya, 2008, "Hindi Urdu machine transliteration using finite-state transducers", *COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics Manchester,* Volume 1.

Redhouse, J. W., 1884, *A simplified grammar of the Ottoman Turkish language*, published by Trübner, London.

Saini, T. S. and G. S. Lehal, 2008, "Shahmukhi to Gurmukhi Transliteration System: A Corpus based Approach", *Natural Language Processing and Applications Research in Computing Science* 33.

Şişman, A. N., 1995, *Ottoman Text Transcription*, M.S. Thesis, Middle East Technical University.

Timutaş, D., 2003, *Ottoman Turkish Grammar, volume 3 [in Turkish]*, 10th edition, Alfa Publishing, İstanbul.

Wells, C., 1880, A Practical Grammar of Turkish language, Published by Bernard Quaricth, London,.

Large Corpora for Turkic Languages and Unsupervised Morphological Analysis

Vít Baisa, Vít Suchomel

Natural Language Processing Centre
Masaryk University, Brno, Czech Republic
{xbaisa, xsuchom2}@fi.muni.cz

Abstract

In this article we describe six new web corpora for Turkish, Azerbaijani, Kazakh, Turkmen, Kyrgyz and Uzbek languages. The data for these corpora was automatically crawled from the web by SpiderLing. Only minimal knowledge of these languages was required to obtain the data in raw form. Corpora are tokenized only since morphological analyzers and disambiguators for these languages are not available (except for Turkish). Subsequent experiment with unsupervised morphological segmentation was carried out on the Turkish corpus. In this experiment we achieved encouraging results. We used data provided for MorphoChallenge competition for the purpose of evaluation.

1. Introduction

Obtaining textual data from the web has become a popular way to build large corpora for linguistic research. All web data is in an electronic form, instantly accessible, in large volume and covering various topics in many languages.

On the other hand, the internet is quite wild: messy, unordered and much duplicate. Solutions to these problems are being developed by other researchers such as (Pomikálek, 2011) whose text cleaning software was used in this work.

Since the performance of NLP generally tends to improve with increasing amount of training data, our aim is to obtain as much grammatical sentences as possible. Many words occur sparsely (according to Zipf's law), so we need really huge text collections to be able to study rare words' behaviour on sufficient number of their utterances.

Turkic languages are interesting for their productive inflectional and derivational agglutinative morphology which causes that these languages have immense amount of various wordforms. Comparing two corpora of the same size: English and Turkish, the second will contain much more wordforms but with lower frequencies. Thus, for these languages, the need for large corpora is even more pronounced.

We chose Turkish, Azerbaijani, Uzbek, Kazakh, Turkmen and Kyrgyz for our work since these languages are more or less connected to the corresponding nations and countries. Unlike other Turkic speaking areas, there are internet top level domains associated with the selected countries. That is why we decided not to collect Uyghur and Tatar texts.

2. Related work

2.1. Building web corpora

Building web corpora has received much attention recently. Table 1 presents selected previous work showing that it is possible to create very large corpora from the web.

The successful techniques used in the former works are search engines querying, web crawling (traversing the internet and downloading documents) and thorough data post-processing. Also, (Baroni et al., 2006) present a web tool able to build a web corpus almost instantly. It performs all necessary steps to prepare the data for further studying, such as concordance queries or terms extraction. However, we argue building billions scale corpora using that tool would require massive search engine querying which could turn out problematic.

We took advice from the previous works and developed new crawler SpiderLing (Suchomel and Pomikálek, 2012). We used the crawler in cooperation with several tools developed by authors referenced in Table 1.

2.2. Corpora of Turkic languages

Probably the largest corpus for Turkish till now was *BOUN Corpus* (Sak et al., 2008) containing about 423M words and 491M tokens. Among others are *METU* corpus with 2M words whose part also forms Turkish METU–Sabanci Treebank (Say et al., 2002), 50M web corpus (Dalkiliç and Çebi, 2002), Turkish part of parallel corpus of Balkan languges containing about 34M tokens (Tyers and Alperen, 2010) and recently developed Turkish corpus with about 42M words containing also Turkish word sketches (Ambati et al., 2012). Still under development is Turkish National Corpus with target size 50M words (Aksan and Aksan, 2009).

Probably the largest corpus for Azerbaijani is described in (Mammadova et al., 2010), containing about 300M words but since there is no mention about boilerplate removing, cleaning and de-duplicating, it is hard to estimate actual size of the corpus.

As for Kazakh, Kyrgyz, Uzbek and Turkmen languages there are some corpora for these langauges but either very small or not accessible (only mentioned in papers, on web pages).

(Biemann et al., 2004) developed corpora of relatively small size for Kazakh, Kyrgyz, Azerbaijani, Turkish, Chuvash, Uzbek, and Tatar mostly from Wikipedia.

2.3. Unsupervised morphology segmentation

Developing corpora of Turkic languages with almost no language tools available, we are forced to use unsupervised methods. Fortunately, unsupervised morphology analysis

Table 1: Overview of selected previous work concerning building large web corpora

language	reference	corpus size
English	(Liu and Curran, 2006)	10 bn tokens
German, Italian	(Baroni and Kilgarriff, 2006)	3.6 bn tokens total
English	(Pomikálek et al., 2009)	5.5 bn words
Dutch, Hindi, Indonesian, Norwegian, Swedish, Telugu, Thai, Vietnamese	(Kilgarriff et al., 2010)	680 mil words total
American Spanish, Arabic, Czech, Japanese, Russian	(Suchomel and Pomikálek, 2012)	51.6 bn tokens total

and segmentation have been well studied since 2000's. Several methods were proposed: (Bernhard, 2006; Demberg, 2007; Snyder and Barzilay, 2008; Argamon et al., 2004) with *Morfessor* (Creutz et al., 2005) being probably the most significant representative of them.

For evaluative purpose, competition *MorphoChallenge* (Kurimo et al., 2006) has been organized several times since 2005 with focus on English, Finnish and Turkish languages. We used their evaluation method for our experiment described in 4.3..

3. Building six Turkic web corpora

There are many ethnic groups and language varieties mixed together in the six language–country pairs we selected. Moreover, some of the languages are somewhat spoken in other countries too. Since we do not understand Turkic languages, we had to carefully constrain crawling and post-processing of the data. Fortunately, most web sites offer documents in just one or two languages understood by major part of the population, but we could not rely on that. Crawling was limited to the respective internet top level domain.

Furthermore, five of the six languages currently use two or three writing systems. We decided to collect the scripts prevailing in the recent texts: Latin for Azerbaijani, Uzbek, Turkmen and Cyrillic (with extensions) for Kazakh and Kyrgyz.

Three language specific models for each selected language were trained using texts from the respective Wikipedia. Byte trigrams for character encoding detection (tool Chared[1]), character trigrams for language identification[2] and a wordlist for boilerplate removal. Filtering crawled documents through these tools/models further helps eliminating unwanted content. Thanks to the strict limits, we believe a good quality of the texts was achieved at the cost of the resulting corpora size.

A couple of seed URLs (the links to start the crawling with) is usually enough in a network of websites densely connected by many links. Since the Turkic presence on the internet is relatively scarce, we obtained more starting URLs to cover more websites (see Table 2) using Corpus Factory (Kilgarriff et al., 2010) and Wikipedia. To get more texts from scarce resources, we configured the crawler to visit websites with less text amount than usually expected.

[1] nlp.fi.muni.cz/projects/chared/

[2] code.activestate.com/recipes/326576

Table 3: Processing the Turkish web. Each line represents a crawling or post-processing step which prevented some data not to pass. Only the last part was put in the final corpus.

data processing phase	fraction of documents	fraction of data size
HTML not retrieved	22.0 %	—
wrong encoding detected	0.7 %	—
other language	17.0 %	13.5 %
boilerplate	14.4 %	49.0 %
exact duplicates	17.8 %	14.7 %
near duplicates	15.9 %	16.4 %
clean text	12.1 %	6.4 %

The texts were tokenized on spaces, punctuation was treated as a separate token. Boilerplate (HTML markup, very short paragraphs and non-grammatical sentences) was removed by Justext[3] (Pomikálek, 2011). Duplicate and near-duplicate paragraphs were removed by n-gram based deduplication tool Onion[4] (Pomikálek, 2011). Misspelling was not dealt with.

Table 2 contains information about data size during crawling and processing. A detailed view on processing the Turkish corpus is presented in Table 3. The corpora have been installed in SketchEngine[5] with enabled concordance querying and wordlist functionality. The final sizes of the corpora in SketchEngine are displayed in Table 4.

4. Unsupervised morphological analysis

4.1. Motivation

Despite there are some morphological analyzers for Turkic languages, namely *TRmorph* (Çöltekin, 2010) and two-level analyzer (Oflazer, 1994) for Turkish, *UZMORPP* for Uzbek (Matlatipov and Vetulani, 2009) and *Azmorph* for Azerbaijani developed within *Apertium* project (Forcada et al., 2009), we are interested in unsupervised methods since other Turkic languages are uncovered in this respect.

A morphological analysis and disambiguation should assign one lemma and one morphological tag to each token in a corpus. With this information one can search for more

[3] code.google.com/p/justext/

[4] code.google.com/p/onion/

[5] the.sketchengine.co.uk

Table 2: Size of crawled HTML data, filtered plaintext and deduplicated texts. $Crawler's\ yield\ rate = \frac{plaintext\ size}{raw\ data\ size}$. $Final\ yield\ rate = \frac{deduplicated\ plaintext\ size}{raw\ data\ size}$.

language	initial domains	raw data [MB]	plaintext [MB]	crawler's yield rate	deduplicated plaintext [MB]	final yield rate	crawling time [h]
Azerbaijani	727	61,479	4,644	7.55 %	834	1.36 %	168
Kazakh	431	68,817	9,425	13.70 %	1,935	2.81 %	168
Kyrgyz	277	13,646	787	5.77 %	271	1.99 %	151
Turkish	157	2,763,780	159,054	5.75 %	26,844	0.97 %	336
Turkmen	51	1,469	113	7.66 %	17	1.18 %	27
Uzbek	454	7,825	497	6.35 %	141	1.80 %	70

Table 4: Turkic corpora obtained using SpiderLing

language	tokens	words	raw wordlist	clean wordlist
Azerbaijani	115M	92M	1.7M	1.4M
Kazakh	175M	136M	2.4M	1.9M
Kyrgyz	24M	19M	684K	590K
Turkish	4,124M	3,370M	20.5M	16.1M
Turkmen	2M	2M	230K	200K
Uzbek	24M	18M	626K	320K

general concordances and e.g. discover grammatical collocates using queries with lemmata and morphological tags.
Although we do not have taggers for several Turkic languages we nevertheless want to provide users with more than just simple querying using regular expressions on wordforms.
As was mentioned, there are some unsupervised methods for morphological analysis (assigning of morphological tags) but we plan to exploit particularly morphological segmentation since this (sub)task of morphological analysis is believed to be much simpler with more reliable results (in the realm of unsupervised methods).
Morphological segmentation splits wordforms into smaller parts: stems, prefixes and suffixes. If we assigned appropriate segmentations to all wordforms in a corpus we would be able to find more general concordances based on queries using stems. In this respect, stems could partially compensate absence of lemmata and tags in a corpus.
The quality of the segmentation is crucial for this enhancement so we evaluated unsupervised segmentations obtained by tool Morfessor-MAP (Creutz et al., 2005). For unsupervised morphological segmentation, Morfessor needs only a wordlist and it was chosen because of its fine results comparing to its competitors and because of being purely unsupervised.

4.2. Evaluation of Morphological Segmentation

For evaluation we used a tool provided for competition *MorphoChallenge* 2005 (Kurimo et al., 2006). Within the competition, gold standards for English, Finnish and Turkish language were provided containing one or more possible morphological segmentations of selected wordforms.

Table 5: Quality of segmentation for various training data.

prec	recall	f-sc	source	WL size
71.15	72.55	71.84	100k	22,3k
77.37	**69.74**	**73.36**	500k	70,4k
72.11	**69.74**	70.90	500k	70,4k
73.83	68.10	70.85	1M	112,6k
73.75	65.09	69.15	5M	313,8k
76.20	65.53	70.46	10M	482,4k
79.90	65.20	65.30	WIN	582,9k
79.10	37.90	51.30	M1	582,9k
73.70	65.10	69.20	M2	582,9k
77.50	65.00	66.40	M3	582,9k

That is why we could evaluate only Turkish segmentations. Nevertheless, we suppose that for other Turkic languages, the quality would be similar.[6]
The evaluation is based on the placement of morpheme boundaries. For example Turkish word *taylanddaki* (in Thailand) should be segmented into two parts: *tayland* and *daki*.[7]
Every correctly placed morpheme boundary forms *hit* (H), missing morpheme boundary forms *insertion* (I) and redundant boundary forms *deletion* (D). Precision is then the number of hits divided by the sum of the number of hits and insertions: $\frac{H}{(H+I)}$, recall is the number of hits divided by the sum of the number of hits and deletions: $\frac{H}{(H+D)}$ and f-score is as usually the harmonic mean of precision and recall.

4.3. Unsupervised Segmentation Results

In Table 5 there are results for various training data (wordlists) extracted from our Turkish corpus.
First three columns stand for precision, recall and f-score as explained before. The fourth column indicates a source for training. The number (in the upper part of the table) means an amount of tokens in a subcorpus from which a wordlist was extracted. The last column contains number

[6] In general, results (f-measure) for English within MorphoChallenge are better than for Finnish and Turkish. Results for Finnish and Turkish are comparable.

[7] In this case, *daki* should be further segmented into two morphemes *da, ki* but for the purpose of querying corpora, the coarse-grained segmentation is good enough.

of wordforms in appropriate wordlist used for unsupervised training.

The lower part of Table 5 shows selected results from MorphoChallenge 2005 for purpose of comparison. WIN stands for highest precision, recall and f-score achieved by various participants. M1–3 stands for evaluation of three variants of Morfessor.

It is clear that we achieved best results in all three measures. Quite surprising is fact that the best score was achieved using relatively small wordlist with about 70,000 of Turkish wordforms. Lower scores for larger wordlists were probably caused by inappropriate setting of one parameter of Morfessor (perplexity threshold) which must be set according to training data size. We run the process with various thresholds and wordlists but did not achieve better results for any of them. Despite, even with larger wordlists, we achieved better results than any participant of MorphoChallenge 2005.

Among other things, we believe that these results support good quality of the Turkish corpus. Training of Morfessor with data provided for MorphoChallenge did not achieve such good results and we suppose it is caused by rather strict language filtering of text data and diversity of language data in our corpus.

5. Conclusion and future work

We have built corpora for six Turkic languages, Turkish with 3.37 bn words being the largest. We believe the corpora are relevant not only due to their size but also with regard to the easiness with which the texts were obtained.

The actual results for morphological segmentation are encouraging but usefulness of unsupervised segmentation for Turkic and other agglutinative languages must be further investigated.

6. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

7. References

Y. Aksan and M. Aksan. 2009. Building a national corpus of turkish: Design and implementation. In *Working Papers in Corpus-based Linguistics and Language Education*, pages 299–310.

Bharat Ram Ambati, Siva Reddy, and Adam Kilgarriff. 2012. Word Sketches for Turkish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Shlomo Argamon, Navot Akiva, Amihood Amir, and Oren Kapah. 2004. Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of Coling 2004*, pages 1058–1064, Geneva, Switzerland, Aug 23–Aug 27. COLING.

M. Baroni and A. Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90. Association for Computational Linguistics.

M. Baroni, A. Kilgarriff, J. Pomikálek, and P. Rychlý. 2006. Webbootcat: a web tool for instant corpora. In *Proceeding of the EuraLex Conference*, pages 123–132.

D. Bernhard. 2006. Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of 2nd Pascal Challenges Workshop*, pages 19–24.

C. Biemann, S. Bordag, G. Heyer, U. Quasthoff, and C. Wolff. 2004. Language-independent methods for compiling monolingual lexical data. *Computational linguistics and intelligent text processing*, pages 217–228.

Ç. Çöltekin. 2010. A freely available morphological analyzer for turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

M. Creutz, K. Lagus, K. Lindén, and S. Virpioja. 2005. Morfessor and hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *In Proceedings of the Second Baltic Conference on Human Language Technologies*.

G. Dalkiliç and Y. Çebi. 2002. A 300 mb turkish corpus and word analysis. *Advances in Information Systems*, pages 205–212.

V. Demberg. 2007. A language-independent unsupervised model for morphological segmentation. *Annual meeting of Association for Computational Linguistics*, 45(1):920.

M.L. Forcada, F.M. Tyers, and G. Ramírez-Sánchez. 2009. The apertium machine translation platform: five years on. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 3–10.

A. Kilgarriff, S. Reddy, J. Pomikálek, and A. Pvs. 2010. A corpus factory for many languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'10, Malta)*.

M. Kurimo, M. Creutz, M. Varjokallio, E. Arisoy, and M. Saraclar. 2006. Unsupervised segmentation of words into morphemes-morpho challenge 2005, an introduction and evaluation report. In *Proceedings of ICSLP*.

V. Liu and J.R. Curran. 2006. Web Text Corpus for Natural Language Processing. *EACL. The Association for Computer Linguistics*.

S. Mammadova, G. Azimova, and A. Fatullayev. 2010. Text corpora and its role in development of the linguistic technologies for the azerbaijani language. In *The Third International Conference Problems of Cybernetics and Informatics*.

G. Matlatipov and Z. Vetulani. 2009. Representation of uzbek morphology in prolog. *Aspects of Natural Language Processing*, pages 83–110.

K. Oflazer. 1994. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9(2):137.

J. Pomikálek, P. Rychlý, and A. Kilgarriff. 2009. Scaling to billion-plus word corpora. *Advances in Computational Linguistics*, 41:3–13.

J. Pomikálek. 2011. *Removing Boilerplate and Duplicate*

Content from Web Corpora. Ph.D. thesis, Masaryk University, Brno.

H. Sak, T. Güngör, and M. Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. *Advances in natural language processing*, pages 417–427.

B. Say, D. Zeyrek, K. Oflazer, and U. Özge. 2002. Development of a corpus and a treebank for present-day written turkish. In *Proceedings of the eleventh international conference of Turkish linguistics*, pages 183–192.

B. Snyder and R. Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. *Proceedings of ACL-08: HLT*, pages 737–745.

V. Suchomel and J. Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the Seventh Web as Corpus Workshop*, Lyon, France, In print.

F. Tyers and M.S. Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Forthcoming in the proceedings of the LREC workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*.

Demonstrative Anaphora in Turkish: A Corpus Based Analysis

Ayışığı B. Sevdik Çallı
Middle East Technical University
Dumlupinar Blv. No:1, 06800, Cankaya, Ankara, Turkey
E-mail: ayisigi@ii.metu.edu.tr

Abstract

This study investigates Turkish demonstrative anaphora including bare demonstrative uses and demonstrative NP uses on a 20K subpart of the METU Turkish Discourse Bank. Antecedents of demonstrative anaphora including abstract object references are identified in 10 texts of approximately 2000 words each. Preliminary analysis shows that for endophoric cases, where the antecedent of the anaphora can be identified in the text, references to concrete object antecedents of the three Turkish demonstratives *bu* ('this), *şu* ('this/that), and *o* ('that) is overall higher than references to abstract object antecedents. However, for the demonstrative *bu* ('this), abstract object reference is almost equal to the concrete object references. The antecedents of the third person singular pronoun *o* ('he/she/it) have also been identified as it is lexically identical to the distal demonstrative *o* ('that), and it was seen that the demonstrative pronoun use occurred as much as the personal pronoun use. The implications of these findings are discussed in terms of Turkish anaphora resolution.

1. Introduction

Expressions in discourse referring to previously introduced objects denoted by nouns, noun phrases, parts of sentences, or text intervals are known as *anaphoric expressions* or more simply *anaphora*. These expressions can be personal pronouns, possessive pronouns, demonstrative pronouns, proper nouns or definite noun phrases. The previously introduced objects are known as the *antecedents* or the *referents* of the anaphora. Anaphora may involve references to both concrete entities, such as those denoted by noun phrases, or abstract entities, such as those introduced into a discourse by constructions like verb phrases (VPs), or whole sentences (Asher, 1993). This last type of anaphora with abstract object referents is also called *discourse deixis* Webber (1998a,b)[1]. Concrete objects include physical entities, individuals and places, whereas abstract objects (AOs) can be ideas, events, etc. In fact, Asher (1993) classifies AOs as *eventualities* and *purely abstract objects.* According to this classification eventualities consist of *events* (i.e. activities, processes, accomplishments, achievements) and *states*; whereas purely abstract objects consist of *fact-like objects* (i.e. possibilities, situations/state of affairs, facts) and *proposition-like objects* (i.e. pure propositions, projective propositions such as questions, commands, and desires). Dipper & Zinsmeister (2011) also add the category *deverbal* under eventualities when annotating German AO anaphora, where deverbal nouns such as 'limitation' are categorized under this label.

The current study aims to be a preliminary work for the annotation and resolution of abstract object demonstrative anaphora in Turkish. In order to gain a general understanding of Turkish demonstrative anaphora, a corpus-based analysis is done on a 20K subpart of the Metu Turkish Discourse Bank (Zeyrek et al, 2010), where all demonstrative anaphora are resolved and identified as either abstract or concrete references. The following section overviews some related work, while sections 3 and 4 explain the research and the results, respectively. In section 5, some conclusions are drawn, followed by directions for future work.

2. Related Work

There have been some previous corpus studies on demonstrative anaphora in English, as well as in some other languages. Corpora studies in English include Eckert & Strube (2000), Byron (2002), Cokal-Karadas (2005), Botley (2006), Hedberg, Gundel & Zacharsky (2007), and Poesio & Artstein (2008). Some other languages for which demonstrative anaphora has been studied on corpora are French, Portuguese (Vieira, Salmon-Alt & Gasperin, 2005), Danish, Italian (Navarretta & Olsen, 2008), Spanish, Catalan (Recasens & Marti, 2009), German (Dipper & Zinsmeister, 2009) and Korean (Lee & Song, 2010). Table 1 gives a comparative overview of these studies, including information on the size of the corpora annotated.

Anaphora studies for Turkish have mostly focused on the anaphoric use personal pronouns or zero anaphora. Initial investigations have been purely linguistic in nature, such as Enc (1986) and Erguvanli-Taylan (1986), where pronominal and zero anaphora in Turkish are investigated. More recent studies have concentrated on computational approaches for the resolution of anaphora. These include Tın and Akman (1994), Yuksel and Bozsahin (2002), Yildirim, Kilicaslan & Aykac (2004), Tufekci and Kilicaslan (2005), Tufekci et al. (2007), Kucuk and Turhan-Yöndem (2007), Yildirim and Kilicaslan (2007), Yildirim, Kilicaslan & Yildiz (2007), Yildirim, Kilicaslan & Yildiz (2009), and Kilicaslan, Guner and Yildirim (2009). The findings in these studies have further been

[1] In this paper such anaphora will be referred to as *abstract object (AO) anaphora*.

Study	*Corpus (size annotated)*	*Type(s) of Anaphora Annotated*
Eckert and Strube (2000)	English Switchboard corpus dialogs (size not available)	pers. & dem. pronouns
Viera et al. (2002)	Portuguese and French (50 dem. NPs each)	dem. NPs
Byron (2002)	English problem-solving dialogs from TRAINS93 corpus (10K)	pers. & dem. pronouns
Cokal-Karadas (2005)	English academic written discourse (586 journal articles)	*this, that*
Botley (2006)	English spoken discourse, news and literature (300K)	*this, that, these, those*
Hedberg, Gundel and Zacharsky (2007)	New York Times newspaper articles (2 full issues, 321 demonstratives)	*that, this*
Poesio and Artstein (2008)	English Arrau Corpus mixed texts (95K)	all NPs and pronouns
Navarretta and Olsen (2008)	Danish texts (60K) and Italian texts (55K)	(zero) pers. & dem. pronouns
Recasens and Marti (2009)	Catalan and Spanish newspaper/newswire articles (400K each)	(zero) pers. & dem. NPs
Dipper and Zinsmeister (2009)	Europarl corpus (32 German texts, ~20 sentences each, 48 instances of *this*)	*this* (Ger. dies)
Lee and Song (2010)	Korean spoken and written corpora (20 K)	dem. pronouns

Table 1: Comparison of Demonstrative Anaphora Studies

applied to some NLP applications, such as in Kucuk and Yazici (2008, 2009) and Can et al. (2008, 2010). All these studies have concentrated on personal pronouns or other pronominal anaphora referring to concrete objects. Only a handful of studies have been conducted specifically on Turkish demonstratives. One such study is Turan (1997), where the use of the demonstratives *bu* ('this) and *şu* ('this/that) on the Bilkent University e-database consisting of newspaper articles and novel texts having a total of 56 demonstratives has been studied. Another study is Cokal-Karadas (2010), which shows the similarities and differences between *bu-şu* and *this-that* in journals of linguistics and language education within the framework of Rhetorical Structure Theory (Mann & Thompson, 1988). To the best of our knowledge, there has been no attempt to systematically observe a corpus of Turkish specifically for demonstrative anaphora referring to abstract objects. The current paper reports the preliminary investigations and results of such a corpus study.

3. The Study

3.1 Basic Information About Turkish Demonstrative Pronouns

Turkish demonstrative pronouns have three main types, given in Lewis (1967) as *bu* 'this', *şu* 'this/that', and *o* 'that'. Göksel and Kerslake (2005) also provides the plural forms as *bunlar* 'these', *şunlar* 'these/those', and *onlar* 'those'.

The main difference between these three pronouns is described as a difference in proximity. In the simplest sense, closer objects are referred with *bu* ('this), farther objects are referred with *şu* and objects that are furthest away are referred with *o*. However, *şu* ('this/that) is often conceived to be accompanied by an ostensive gesture of pointing. Göksel and Kerslake (2005) also state that *şu* ('this/that) implies that the referent is newly introduced, whereas *bu* ('this) does not, and they cannot be substituted for each other. In addition to this, the referent of *şu* ('this/that) may succeed it after a colon. In cases where a previously mentioned concrete item that is out of sight for both the speaker and the hearer is referred to, *o* ('that) is used. If an object in context is to be topicalized, then either *bu* or *o* can be used.

3.2 The Corpus

In order to obtain a preliminary view of demonstrative anaphora in Turkish, the first 10 texts of approximately 2000 words each in Subcorpus1[2] of the METU Turkish Discourse Bank (TDB) (Zeyrek et al., 2010) was analyzed. The portion of Subcorpus1 analyzed for this study is approximately 20K-words consisting of texts from the genre "novel".

3.3 Method

For this study, all uses of Turkish demonstrative pronouns (i.e. *bu* ('this) /*şu* ('this/that) /*o* ('that)), including bare demonstrative usages and demonstrative+NP usages, have been identified with their antecedents. As the future goal of this preliminary work is to eventually study abstract object anaphora for Turkish and provide a computational resolution method, the antecedents have also been identified as being abstract, concrete objects, or exophoric references to text-external material. In (1) reference to an abstract object is shown, where *bunun* ('this+Gen) refers to *getting sick*. (2) exemplifies a concrete reference to *his father's study* by the demonstrative+NP *this room*, where as the referent for *those devious pains* in (3) cannot be found in the text, hence it is marked as exophoric.

(1) Şemsî Ahmed Paşa onu ayakta karşılayarak, başına gelen talihsiz kazaya çok üzüldüğünü, eğer <u>*hasta olursa*</u> **bunun** sorumluluğunun kendisinde olduğunu söyledi.
'Şemsî Ahmed Pasha greeting him on his feet,

[2] Subcorpus1 is one of four 400K-word subcorpora of the Metu Turkish Corpus (MTC) (Say et al., 2004) prepared as part of the TDB project, which retains the same genre distribution of the main MTC corpus.

said he was very sorry for the unfortunate accident that happened (to him), if he were to *get sick* the responsibility of **this** would be on himself.

(2) Üst katta *babasının çalışma odasına* girdi. Onun ölümünden beri ilk kez girdiği **bu odayı** ağır bir koku kaplamıştı.
'He entered *his father's study* upstairs. A heavy stench had filled **this room**, which he entered for the first time after his death.

(3) Her yere kendisiyle birlikte taşımaz mı içindeki **o sinsi acıları**?
'Does he not carry **those devious pains** with himself everywhere?

(4) Gözlerine bakamazdım ben *insanların*. Korkaktım ben. Ben **onlardan** korkardım, kızgınken bile.
'I couldn't look into *people*'s eyes. I was a coward. I was afraid of **them**, even when I was angry.

Apart from the demonstratives, explicit third person pronouns have also been identified and resolved. The reason for identifying the third person pronouns is that in Turkish the demonstrative *o* ('that) is homonymous to the third person pronoun. In order to find distinguishing features for these two uses, third person pronouns have also been resolved as exemplified in (4).
All analysis was done manually by a single annotator (the author) for this preliminary work.

4. Results

A total of 682 instances of demonstrative anaphora were identified in the 20K portion of the TDB. Usages of *bu*, *şu*, *o* as abstract anaphora was identified in 131 cases versus 224 concrete anaphora uses, 126 references by personal pronouns and 201 of the cases were identified as exophoric uses, where the referent was not mentioned in the text.
AO anaphora for the demonstrative pronoun *bu* dominated the results (106 cases), where there were only two cases of *şu* anaphora, and 23 cases of *o* anaphora. On the other hand, 224 concrete object referents of demonstratives were distributed as: *bu* (135), *şu* (2), *o* (87). Apart from the 110 demonstrative uses of the pronoun *o*, there were 126 personal pronoun uses. Finally, 201 of the cases observed were found to be exophoric. (See Table 2).
If the exophoric and personal pronoun cases are excluded, which make up about 48 percent of all the anaphoric cases, the remaining 355 instances of demonstrative anaphora have the following distribution: about 37% includes references to abstract entities, whereas the remaining 63% refers to concrete entities. 138 of these anaphora are pure demonstrative anaphora, i.e. referencing by bare demonstrative uses, without NP complements as in (1); the rest are demonstrative NP uses as in (2) and (3). 75% of the bare uses involve the demonstrative *bu*, 24% involve *o* and only 1% is using *şu*. Within the pure demonstrative cases, reference to abstract and concrete objects are nearly equally distributed (i.e. 68 and 70 instances, respectively). The abstract references occur

	Abstract	***Concrete***	***Pers. Prn.***	***Exophoric***	***Total***
bu	61	43	0	11	115
bu+NP	45	92	0	67	204
şu	1	0	0	0	1
şu+NP	1	2	0	21	24
o	6	27	126	47	206
o+NP	17	60	0	55	132
Total	131	224	126	201	**682**

Table 2: Distribution of Turkish Demonstrative Anaphora

more with the use of the demonstrative *bu*, i.e. 61 abstract uses versus 43 concrete uses (see Figures 1 and 2). On the other hand, concrete references occur more with the demonstrative *o*, i.e. 27 concrete uses versus 6 abstract uses. Distribution of *bu*, *şu,* and *o* are similar in demonstrative NP uses also, i.e. 63%, 1% and 36%, respectively. However, it can be said that demonstrative NP anaphora favors concrete referents, as they are more than twice as much as abstract referents (i.e. 154 concrete cases versus 63 abstract cases). The *bu+NP* uses dominate references to both concrete (*bu*:92, *şu*:2 , *o*:60) and abstract objects (*bu*:45 , *şu*:1 , *o*:17).
18.48% of all anaphora was made up of cases involving the 3rd person pronoun "o" (126 cases). These included references to proper nouns as in (5), as well as reference to NPs as in (6).

(5) Ermeni Ante, *Mihriban Hanım*'a duyduğu aşkı yol boyu taşıdı. Çarpışma sürerken bile **onun** yüzünü görüyordu.
'Armenian Ante carried his love for *Lady Mihriban* through the journey. Even during the battle (he) would see **her** face.

(6) Benimle aynı saatlerde bazen bir *kör adam* da biniyordu metroya. **Ona** birkaç kez rastlamıştım.
'Sometimes *a blind man* also rode the subway at the same time as me. (I) had come across **him** a few times.

201 of the cases had exophoric referents. Most of these anaphora were simply just unmentioned in text, such as the case of (7), referring to the particular day of the event described. Some included ostension as in (8), whereas some were specialized uses (28 cases) as in the use of "*o kadar*" ('so much) in (9). Other special uses involved "bu kadar" ('this much) (9 cases) and "o zaman" ('that time) (6 cases). Some others were vague references via a personal pronoun (47 cases) to an unmentioned salient person.

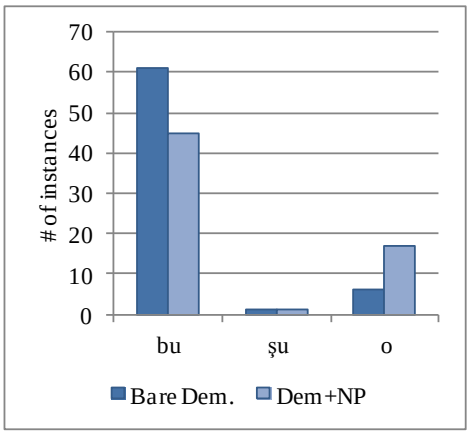

Figure 1: Distribution of Demonstrative Anaphora with Abstract Object Referents

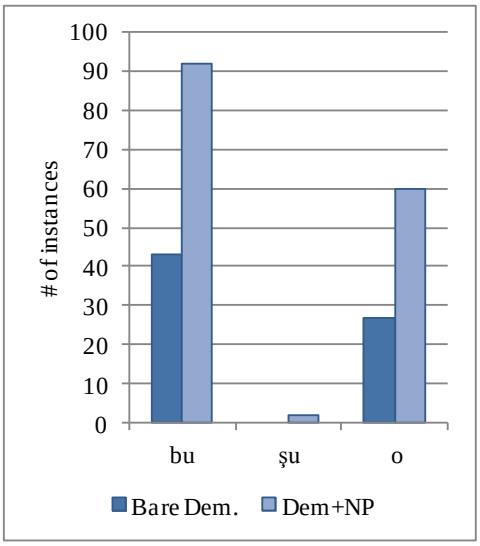

Figure 2: Distribution of Demonstrative Anaphora with Concrete Object Referents

(7) Sabah çok erken saatte bir önceki akşam gün batmadan hemen önce astığı çamaşırları toplamaya çıkıyordu ve doğal olarak da gün batmadan **o günkü** çamaşırları asmak için geliyordu.
'(She) would go out very early in the morning to collect the laundry (she'd) hung up the previous evening just before the sun set and naturally would arrive before sun set to hang up the laundry for **that day**.

(8) -"Çek **şu lambayı** gözümden. Sen kimsin?"
'Put **this/that lamp** away from me. Who are you?

(9) Size kendi hayatımdan anılar anlatacağım. **O kadar** çok şey hatırlıyorum ki...
'I'm going to tell you memories from my life. I remember **so (much)** many things.

There were also a total of 11 cataphoric instances observed, where the referent was found after the anaphor (*bu*:3, *şu*:2, *o*:6). All, except one of these involved reference to concrete objects, 5 of which were personal pronoun anaphora. In example (10), a cataphoric use of *bu* is given, where the antecedent succeeds the anaphor. The only cataphoric abstract reference observed was by *şu*.

(10) Gece de yatmadan önce, arka bahçeye - bahçe değil **bu**, *beton döşeli bir aralık*tı- bakan yatak odası penceresinin perdelerini açmamıştım.
'Before going to bed at night, I had not opened the curtains of the bedroom window overlooking the backyard – **this** is not a yard, it's *a concrete paved gap*.

5. Conclusions and Future Work

Some preliminary conclusions can be drawn about Turkish demonstrative anaphora observed in novels. First of all, it is observed that about 1/3rd of all the demonstrative anaphora in Turkish novels consists of exophora, whereas the rest is endophoric (i.e. within text) uses. Within the endophora, the most frequently used demonstrative in Turkish is found to be *bu* and it is also the most preferred demonstrative for AO reference, where *şu* and *o* are rarely used. On the other hand, demonstrative use of *o*, is preferred for referencing concrete objects. There is also substantial use of *o* as a personal pronoun, dominating all its other uses in terms of frequency.

Ongoing work involves the annotation of this data by a second annotator, which will provide a means to calculate agreement statistics and ensure the reliability of the results obtained here for future computational analyses. This also includes the clarification of annotation guidelines for annotating Turkish demonstrative anaphora. Future work involves the annotation of semantic types of the referents, especially aiming to identify a degree of abstractness of the referents as determined by Asher (1993). Further work will increase the size of the corpus data observed, as well as including other genres.

As AO anaphora in Turkish is not a well investigated topic, the ultimate goal of the project is to identify the distinguishing features for abstract object demonstrative anaphora in Turkish and to develop a computational resolution algorithm for such anaphora. Current results suggest some implications for such future work. First of all, concentrating on bare demonstrative uses would eliminate most of the exophoric uses, which are found to be more frequent in demonstrative NP uses. Another reason may be that demonstrative NP uses clearly favor concrete references, whereas bare uses, especially for the more frequent demonstrative *bu* favor reference to AOs. One other consideration for automatic anaphora resolution work can be to distinguish and exclude personal pronoun uses of *o*.

6. Acknowledgement

I would like to thank Deniz Zeyrek for her valuable comments and contributions.

7. References

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrehct: Kluwer Academic Publishers.

Botley, S.P. (2006). Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1), pp. 73--112.

Byron, D. (2002). Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 80--87.

Can, F., Koçberber, S., Balçik, E., Kaynak, C., Ocalan, H.C., and Vursavas, O.M. (2008). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*, 59(3), pp. 407--421.

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C. and Uyar, E. (2010). New event detection and topic tracking in Turkish. *Journal of the American Society for Information Science and Technology*, 61(4), pp. 802--819.

Cokal-Karadas, D. (2005). A contrastive analysis of the pronominal usages of this and that in academic written discourse. MA Thesis, Department of English Language Teaching, Middle East Technical University, Ankara, Turkey.

Cokal-Karadas, D. (2010). The pronominal bu-şu and this-that: rhetorical structure theory. *Dilbilim Araştırmaları (Linguistics Research)*, 1.

Dipper, S. and Zinsmeister, H. (2009). Annotating discourse anaphora. In *Proceedings of the Linguistic Annotation Workshop III*, pp. 166--169.

Dipper, S. and Zinsmeister, H. (2011). Annotating abstract anaphora. *Language Resources & Evaluation*, Online First, 3. September 2011.

Eckert, M. and Strube, M. 2000. Dialogue acts, synchronizing units and anaphora resolution. *Journal of Semantics*, 17(1), pp. 51--89.

Enc, M. (1986). Topic switching and pronominal subjects in Turkish. In D. Slobin and K. Zimmer (Eds.), *Studies in Turkish Linguistics*. Amsterdam: John Benjamins, pp. 195--208.

Erguvanli-Taylan, E. (1986). Pronominal versus zero representation of anaphora in Turkish. In D. Slobin and K. Zimmer (Eds.), *Studies in Turkish Linguistics*. Amsterdam: John Benjamins, pp. 209--233.

Göksel, A. and Kerslake, C. (2005). *Turkish: A comprehensive grammar*. London: Routledge.

Hedberg, N., Gundel, J.K. and Zacharski, R. (2007). Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In *Proceedings of DAARC-2007*, pp. 31--36.

Kilicaslan, Y., Güner, E.S. and Yildirim, S. (2009). Learning-based pronoun resolution for Turkish with a comparative evaluation. *Computer Speech and Language*, 23(3), pp. 311--331.

Kucuk, D. and Turhan-Yondem, M. (2007). A knowledge-poor pronoun resolution system for Turkish. In *Proceedings of 22nd International Symposium on Computer and Information Sciences (ISCIS 2007)*, pp. 1--6.

Kucuk, D. and Yazici, A. (2008). Identification of Coreferential Chains in Video Texts for Semantic Annotation of News Videos. In *Proceedings of 23rd International Symposium on Computer and Information Sciences (ISCIS 2008)*, pp.1--6.

Küçük, D. and Yazici, A. (2009). Employing Named Entities for Semantic Retrieval of News Videos in Turkish. In *Proceedings of the 24th International Symposium on Computer and Information Sciences (ISCIS 2009)*, pp. 153--158.

Lee, S. and Song, J. (2010). Annotating Korean demonstratives. In *Proceedings of the Fourth Linguistics Annotation Workshop.* pp. 162--165.

Lewis, G. (1967). *Turkish Grammar*. London: Oxford University Press.

Mann, W.C. and Thompson, S.A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8, pp. 244--277.

Navaretta, C. and Olsen, S. (2008). Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC 2008*, pp. 2046--2052.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *the Proceedings of the LREC Workshop on Language Resource and Language Technology Standards* (LREC 2008), pp. 1170—1174.

Recasens, M. and Marti, M.A.. (2009). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources & Evaluation*. 44(4), pp. 315--345.

Say, B.; Zeyrek, D.; Oflazer, K. and Ozge, U. (2004). Development of a corpus and a treebank for present-day written Turkish. In K. Imer and G. Dogan (Eds.), *Current Research in Turkish Linguistics: Proceedings of the 11th International Conference on Turkish Linguistics.* Eastern Mediterrean University Press, pp. 183--192.

Tin, E. and Akman, V. (1994). Situated processing of pronominal anaphora. In *Proceedings of Second Conference for Natural Language Processing (KONVENS'94)*. University of Vienna, Austria, pp. 369--378.

Turan, U.D. (1997). Metin işaret adılları: *Bu, şu* ve metin yapısı (Discourse Deictic pronouns *this, that*, and the structure of discourse). In D. Zeyrek & Ş. Ruhi (Eds.), *XI. Dilbilim Kurultayı Bildiriler (The Proceedings of the XI. Linguistics Conference)*. Ankara: Middle East Technical University, pp. 201--212.

Tufekci, P. and Kilicaslan, Y. (2005). A computational model for resolving pronominal anaphora in Turkish using Hobbs' naïve algorithm. In *Proceedings of World Academy of Science, Engineering and Technology (WEC)*, pp. 13--17.

Tufekci, P., Kucuk, D., Turhan-Yöndem, M. and Kilicaslan, Y. (2007). Comparison of a syntax-based and a knowledge-poor pronoun resolution systems for Turkish. In *Proceedings of the International*

Symposium on Computer and Information Sciences (ISCIS 2007). Middle East Technical University, p.53.

Vieira, R., Salmon-Alt, S., Gasperin, C., Schang, E. and Othero, G. (2005). Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In A. Branco, T. McEnery, and R. Mitkov (Eds), *Anaphora Processing: Linguistic, Cognitive and Computational Modeling.* Amsterdam: John Benjamins, pp. 385--401.

Webber, B.L. (1988a). Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics (ACL)*. Buffalo, New York, pp. 113--122.

Webber, B.L. (1988b). Discourse deixis and discourse processing. Technical Report No. MS-CIS-88-75, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania.

Yildirim, S., Kilicaslan, Y., Aykaç, R.E. (2004). A computational model for anaphora resolution in Turkish via centering theory: an initial approach. In *Proceedings of International Conference on Computational Intelligence*, pp. 124-–128.

Yildirim, S. and Kilicaslan, Y. (2007). A machine learning approach to personal pronoun resolution in Turkish. In *Proceedings of 20th International FLAIRS Conference, FLAIRS-2007*. Key West, Florida, pp. 269--270.

Yildirim, S., Kilicaslan, Y. and Yildiz, T. (2007). A decision tree and rule-based learning model for anaphora resolution in Turkish. In *Proceedings of the 3rd Language and Technology Conference (LTC'07)*. Poznañ, Poland, pp. 89--92.

Yildirim, S., Kilicaslan, Y. and Yildiz, T. (2009). Pronoun resolution in Turkish using decision tree and rule-based learning algorithms. In Z. Vetulani and H. Uszkoreit (Eds.), *Human Language Technology: Challenges of the Information Society, Proceedings of the Third Language and Technology Conference, LTC 2007*. Berlin-Heidelberg: Springer-Verlag, pp. 270--278.

Yuksel, O. and Bozsahin, C. (2002). Contextually appropriate reference generation. *Natural Language Engineering*, 8(1), pp. 69--89.

Zeyrek, D., Demirsahin, I., Sevdik-Calli, A.B., Ogel-Balaban, H., Yalcinkaya, I., and Turan. U.D. (2010). The annotation scheme of the Turkish discourse bank and an evaluation of inconsistent annotations. In *Proceedings of theFourth Linguistic Annotation Workshop*(*LAW IV)*, pp. 282--289.

Towards a Morphological Annotation of the Khakass Corpus

Alexandra V. Sheymovich
Institite of Lingustics Russian Academy of Science (RAS), Moscow
asheimovich@yandex.ru
Anna V. Dybo
Institite of Lingustics Russian Academy of Science (RAS), Moscow
adybo@mail.ru

Abstract

This paper describes development of a corpus of the Khakass language and design of a morphological parser for it. Being one of the RAS projects, it follows the RAS program in regard to the development of corpora for languages of the Russian Federation, including Turkic minority languages such as Khakass. Khakass is a language spoken by about 20,000 people, most of whom are bilingual in Russian. They live in the southern Siberian Khakass Republic in Russia. We present the preliminary linguistic work for creating automatic morphological annotation for the Khakass written corpus. Main components of this work are: 1) the database of the Khakass word stems generated by StarLing system, 2) the computational model of a Khakass wordform and 3) the set of phonetic rules that constrain the choice of allomorphs within the wordform. We also present Khakass inflectional affixes with their allomorphs.

Keywords: corpus of a language, morphological annotation, morphological parser, inflection, computational model of a wordform

1. Background

Khakass is a Turkic language spoken by about 20,000 (of 75,000) Khakass people, most of whom are bilingual in Russian. Most of them live in south-central Siberia in the Khakass Republic in Russia, with the capital city of Abakan [1]. A.Dybo and O.Mudrak classify Khakass as belonging to the Khakass-Altaic group of the Eastern brunch of Turkic languages [2]. While Khakass is written in the Cyrillic alphabet, we transliterated the examples in this paper into Latin alphabet.

This paper describes the preliminary linguistic work to prepare input for an automatic morphological analyzer. The work follows the framework of the RAS corporate project in regards to the development of corpora for languages of the Russian Federation, including Turkic minority languages such as the Khakass. For details please see [3].

In recent years there have been multiple proposals developing recourses for machine processing of Turkic languages, such as spell-checkers and automatic morphological analysers. Examples include Turkish (Çöltekin 2010), Kazakh [3], Azerbaijani [5], Bashkir [6]. The Khakass language has never had any corpora or related machine processing resources, so today we can hardly speak of large samples of its written texts. At present the corpus for the Khakass Language is under development. It includes some of the original Khakass prose translated phrase-by-phrase into Russian and newspaper articles converted into a standardized format. We also used for our corpus the electronic version of the Large Khakass-Russian Dictionary (2006) with its illustrative materials converted in the format of StarLing database (about StarLing see [7]).

2. Motivation for Morphological Annotation of the Khakass Corpus

The effectiveness of a corpus depends on its linguistic annotation. We decided to implement morphological annotation first. It consists of matching of tokens to types (in linguistic terminology, assignment of a wordform to its lexeme), defining classes of stems that combine with some affix, etc.

Morphological analysis tags wordforms with the following grammatical information:

1. Lexeme to which a wordform belongs (a dictionary form of the lexeme);
2. A set of the wordform's grammtical features, known as inflections (for example, case for nouns and tense for verbs);
3. Information about non-standard forms of the wordform, orthographic variations, etc. (Lyashevskaya et al., 2005, 114).

2.1. Main Components of the Morphological Analyser

Three basic components of the morphological analyzer are:

- A stem dictionary that includes phonetic alternation within the stem;
- A computational model of wordform based on an appropriate grammatical description;
- A set of phonetic rules.

2.2. Two-Step Morphological Analysis

We perform morphological analysis in two stages. Initially the tagger detects inflectional morphology, marking grammatical categories such as case, person, tense, etc. Next step is stem lookup in the online dictionary where lexical entries include information about their component derivational affixes (agent suffix, diminutives, etc.).

The motivation behind our approach is two-fold. On the one hand, semantics of the derived words is not fully compositional, i.e. the meaning of a word cannot be predicted completely from the root and affixes' meanings:
(1) *xas* 'river-bank' (noun) – *xasta* 'keep along smth.' (verb):
xana xastərγa[1] 'walk along the fence'
xastərγa < *xas* + *ta* (denominative verbal suffixe) + *-arγa* (suffixe of infinitive)

On the other hand, our formal model of morphology follows Gleason's description of Turkic morphology,

[1] The symbol *ə* here exprsses a front vowel close to *i*, neutral in respect of vowel harmony.

which is discussed in secion 5 below (Gleason 1955). While this model appears adequate for inflectional morphology derivational affixes in Khakass violate its combinatorial tenet prohibiting affixes of the same category from occurring more than once within the same wordform. An example of such violation is the use of the voice affixes in Khackas and other Turkic languages (cf. double causative, combinations of causative and reciprocal, passive and reciprocal, etc.):

(2) *tur*(to stand)-*ɣyz*(Caus)-*ys*(Rec) 'help smb. to put smth.'

Others derivational affixes also co-occur within the same wordform in different ways:

(3) *čük* 'burden' + *te*(verbal aff.)+*n*(Refl) > *čükten* 'to carry' + *žik* (Dimin) > *čük-te-n-žik* 'small sack'

Given the above reasons, we currently are tagging the Khakass corpus only with the inflectional information, but eventually it will incorporate the derivational annotation as well (see Section 4).

Another important feature is that we treat parts of speech such as nouns and verbs as a syntactic, not morphological, distinction. We discuss this point in section 6.

3. Characteristic Features of the Khakass Language Relevant for Morphological Annotation

The following features of Khakass are common to agglutinative languages:

– a developed system of affixes, most of which are grammatically unambiguous (an affix has a single grammatical meaning); grammatical homonymy is uncommon.

– lack of morphological distinction between noun or verb classes (declension/conjugation), i.e. single type of word-altering (cf. inflexional languages);

– lack of significant phonetic alternations in the stems; allomorphs conform strictly to the phonetic rules;

Thus, an agglutinative wordform is constructed by adding to the stem unambiguous standard affixes in a fixed order; morpheme boundaries are distinct; sandhi at boundaries conform to strict rules (see section 7). But these design advantages (desirable from the perspective of the developer of an automatic analyser) are offset by the complexities resulting from the plethora of afore-mentioned sandhi and a very complex paradigm of a wordform due to a large number of its affixes. This is likely to impact performance of a morphological analyzer, which must take into consideration all morpheme combinations permitted in Khakass.

4. Stem Dictionary

The Stem dictionary is automatically extracted from The Large Khakass-Russian Dictionary (2006) using the StarLing database processing system (see [7]). The Stem dictionary is represented as the formatted database including autonomous words in their initial form (lemmas) with all the alternations within the stem which are not predictable from the initial form. If the initial form consists not only of a root morpheme but includes some derivational affixes we place these affixes with their meanings in special fields of the database to allow for the derivational annotation of the corpus in the future. The inventory of derivational affixes can also be used for the future syntactic annotation, as the last affix of the stem allows to define the syntactic function of the wordform.

The process of importing a text file in StarLing and the step-by-step creation of multilevel lexico-grammatical database is described in Krylov (2008).

5. The Model of a Khakass Wordform

The algorithm of automatic morphological annotation is based on a computational model of the wordform. To design such a model we searched for appropriate grammatical description of the Khakass language.

Neither of the available two Khakass grammars such as Baskakov (1953) and Khakass Grammar (1975), in our opinion, provide an adequate description of the Khakass inflection and morphotactics or a level of detail of phonetic changes within a wordform to support automatic morphological analysis. Therefore, we involved, in addition to these sources, elements of the combinatorial grammar defined in Gleason (1955). Gleason classified Turkish morphemes into groups known as orders. Orders were assigned numbers to signify their proximity to the root. "Order 1 consists of all those suffixes which can occur only immediately after the root. Order 2 consists of those which can occur immediately after a morpheme of order 1, or immediately after the root if no morpheme or order 1 is present, but never farther from the root that this. Order 3 consists of those which can occur only after roots or members of orders 1 or 2." It follows that morphemes from the same order cannot co-occur within the same word because then they would have to follow one another and thus belong to different orders by definition. "Orders are, therefore, mutually exclusive classes of morphemes occupying definable places in the sequence of morphemes forming a word" (p. 112).

While Gleason illustrated this concept on Turkish examples, linguists applied his generalization to Turkic and other agglutinative languages mainly spoken in Russia (Mal'tseva (2004, pp. 7–9), Volodin, Khrakovsky (1975), Revzin, Yuldasheva (1969)). The combinatorial morphological grammar is intended to describe languages which meet following requirements: a) fixed order of affixes; b) one-to-one correspondence between the affix and its grammatical meaning; c) the affix of a certain grammeme can appear within the same wordform no more than once. In our work we adopt this morpheme orders as appropriate for Khakass.

We model a wordform as a stem, which adds on a sequence of inflectional affixes. As at present stage of the work we are concerned only with inflection, but not derivation of the corpus, all derivational affixes are included in the stem (e.g., voice of a verb, agent affix of a noun, etc.).

For example, in the word *palyxčylarybystyŋ* 'our fishers'(Gen) affix *-čy* is derivational. It is the agent affix, which forms a word *palyxčy* 'fisher' from *palyx* 'fish'. This affix is recorded in the dictionary, it is considered by the parser as a part of stem and is not included in the morphological analysis:

(4) *palyxčy-lar-ybys-tyŋ*
fisher - Pl. - Poss.1 Pl.- Gen.

Grammatical meaning is assigned to each affix and to all its allomorphs (e.g., *lar/ler/nar/ner/tar/ter* – Plural, *da/de/ta/te* – Locative case, *byn/bin/pyn/pin* –1st Person,

Singular …). For the complete set of Khakass inflectional affixes see Table 1.

6. Inflectional Classes[2] in Khakass

From the prior experience in Altaic studies, we expected to find in Khakass at least three grammatical classes: nouns, verbs and invariables. The first two classes should be characterized by different sets of grammatical features, each of them expressed by a unique formal marker. Each set of formal markers (i.e. affixes) defines the type of inflection and grammatical class of a word, for example, case and number for nouns; person, number, tense for verbs. But even traditional sources such as Baskakov (1953) and Khakass Grammar (1975) say that the difference between grammatical classes is minor. In particular, in the category of words typically classified as nouns, the same word may be a noun, adjective, or adverb depending on its syntactic function: *kičigler ojnapčalar* - 'little ones (children) are playing', *kičig aal* - 'small village', *mənіŋ tasym* - 'my stone', *tas turalar* -'houses of stone' (cf. in English 'stone': *my stone* (noun), *stone wall* (adj), *they stone adulterers to death* (verb)). As an attribute the word becomes invariable, as a subject or object it acquires the features of a noun.

Verbs, nouns and invariables may exhibit the same morphological behaviour. For example, Turkic nouns may accept two sets of person-number affixes – possessive and properly personal: *mənіŋ xol-ym* 'my hand', *mənіŋ alγan-ym* 'my taking' (*ym* – possessive affix 1Sg.); *pis xakas-pys* 'we are Khakasses' (*pys* – 1 Pl.). Verbs have participles, which are nominal forms, varying also in number, case and possessivity. In Khakass there also exists a so-called "Altaic type of the compound sentence" where secondary predication is expressed by case forms of a participle:

(5) *xajdə toγyn-yp*
how work-Conv1
ügren-gle-p-četken-ner-in
study(Refl)-Distr-Form-Prs.Pt-Pl-Poss3-Acc
čooxty- pər-eŋer
tell-Conv1 give-Imp.2Pl
'Tell how do they study, working?' *literally*: 'Tell them-**studying** while working' (Mal'tseva 2004)

It is obvious that words traditionally placed in different grammatical classes receive the same grammatical markers and vary across the same grammatical categories in Khakass. Therefore morphological analyzer can no longer differentiate them by their surface form and must treat these words as belonging to the same grammatical class based on their single inflectional type.

Because the same grammatical markers and categories apply to words traditionally in different classes, annotation cannot encode this distinction for Khakass at the level of morphology, and that's why we treat conventional grammatical classes (and parts of speech) in Khakass on a par in developing an automatic morphological parser. Thus, we developed a single general scheme of the Khakass wordform, which is both verbal and nominal. It is presented in the tabular form (see Table 1).

[2] We are not addressing a controversial for turkologists question of the parts of speech in Khakass because it is largely irrelevant for the automatic morphological analyzer of Khakass.

7. Phonetic Rules of the Khakass Language Relevant for Morphological Annotation

It's important to note that in the present work we don't consider the rules that do not manifest themselves orthographically, as the morphological annotation is meant for processing of written texts.

Due to space limitations we mention here only a few of the main phonetic regularities in Khakass.

1) Vowel harmony

The law of vowel harmony is common for the most agglutinative languages. It states that the quality of the root vowels determines the quality of the following ones in affixes. For Khakass it supposes that words may not contain both front and back vowels. Therefore, most grammatical affixes come in front and back variants. Vowel frontness also affects consonantal place of articulation with uvular *x, γ* following back vowels, velar *k, g* following front vowels: *xarax-tar-ybys-ta* 'in our eyes', *kərek-ten-deŋer* 'on business';

2) Voicing of voiceless consonants in the intervocalic position (*at / ady*);

3) Consonant deletion in the intervocalic position (with vowel lengthening): *xarax* 'eye'+ *ym* > *xaraam* 'my eye';

4) Deletion of narrow vowels (*u, y, i*) of multisyllabic stems before a possessive affix *i/y* (*purun* 'nose' – *purny* < *puruny* 'his nose');

5) Progressive and regressive consonant assimilation (*xus* 'bird' + *lar* > *xus-tar*).

The laws of vowel harmony and assimilation determine the rules for the choice of allomorphs (phonetic variants of inflectional affixes). Afterwards the internal sandhi apply at the morpheme boundaries. The final appearance of a wordform is a result of applying in turn the following two types of rules:

The first group consists of "the rules of choice" for all phonetic variants of affixes. They are presented in tables, as, for example, Table 2 (for affixes of Plural).

We formulated similar rules for affixes of Attribute, Possessive Attribute, Circumstantial Modifier, Person, Emphatic, Possession, Number, Comitative, Negation, Distributive, Subjunctive Mood, Conjunctive Mood, Durative, Perfective, Prospective, Formative, Present Time, Past Time, Future Time, Converb, Optative, Imperative.

The second group of phonetic rules are the so-called "surface rules" describing the mechanism of action of internal sandhi, such as

– voicing of voiceless consonants in the intervocalic position (*at* + *y* > *ady*);

– consonant deletion in the intervocalic position and contraction of two vowels into one long vowel (*xarax* + *ym* > *xaraam, ujγu* + *-γa* > *ujγaa, maŋ* + *-y* > *maa*);

– deletion of a geminate consonant *-γ, -g, -ŋ* (*suγ*+*-γa* > *suγ-a, kög*+*-ge* > *kög-e*);

– deletion of narrow vowels (*u, y, i*) in multisyllabic stems before a possessive affix *i/y* (e.g. *purun* 'nose' – *purny* < *puruny* 'his nose').

Table 1: Model of an inflectional wordform in Khakass

№	0	1	2	3	4	5	6	7	8	9	10	11	12	13		14	15	16	17
	R (S)	Distr	Form	Emph	Perf/ Prosp	Dur	Neg	Tense (Pres Past, Future, Conv), Mood	Irr	Comit	Num (Pl)	Poss	APos	Case		Attr	Emph	Person (1, 2)	Adv
														Simple declension	Possessive declension				
1.		*Γla*	*yp*	*daa*	Perf *ybys*	Dur *ča*	*ba*	Pres *ča*	*ǯyx*	*lyŋ*	*lar*	1sg *m*	*nə*	Gen *nyŋ*	Gen *nyŋ*	*xy*	*ox*	1sg *myn*	Manner *nə*
2.		*Xla*	*ip*	*taa*	Perf *ibis*	Dur *če*	*be*	Pres *če*	*ǯix*	*lig*	*ler*	1sg *ym*	*tə*	Gen *niŋ*	Gen *niŋ*	*ki*	*ök*	1sg *min*	Manner *tə*
3.		*Gle*	*p*	*dee*	Perf *bys*	Dur *čadyr*	*pa*	Pres *čadyr*	*čyx*	*nyγ*	*tar*	1sg *am*	*lə*	Gen *tyŋ*	Gen *tyŋ*	*γy*		1sg *byn*	Manner *lə*
4.		*Kle*	*b*	*tee*	Perf *bis*	Dur *čedir*	*pe*	Pres *čedir*	*čix*	*nig*	*ter*	1sg *um*		Gen *tiŋ*	Gen *tiŋ*	*gi*		1sg *bin*	Temp *n*
5.			*m*	*la*	Prosp *ax*	Dur *čat*	*ma*	Pres *čat*		*tyγ*	*nar*	1sg *öm*		Dat *a*	Dat *a*	*x*		1sg *pyn*	
6.			Form. Neg *bən*	*le*	Prosp *ek*	Dur *čet*	*me*	Pres *čet*		*tig*	*ner*	1sg *em*		Dat *e*	Dat *e*	*k*		1sg *pin*	
7.			Form. Neg *pən*	*na*	Prosp *x*	Dur 1 *ər* (for *par-* and *kəl-* only)		Pres *dyr*				1sg *im*		Dat *γa*	Dat *γa*	*γ*		2sg *zyŋ*	
8.			Form. Neg *mən*	*ne*	Prosp *k*	Dur 1 *ə* (for *par-* and *kəl-* only)		Pres *dir*				1sg *əm*		Dat *ge*	Dat *ge*	*g*		2sg *ziŋ*	
9.				*ox*				Pres *čadadyr*				2sg *ŋ*		Dat *xa*	Dat *xa*			2sg *syŋ*	
10.				*ök*				Pres *čededir*				2sg *yŋ*		Dat *ke*	Dat *ke*			2sg *siŋ*	
11.								Fut *ar*				2sg *aŋ*		Dat *γ* (+*ox*)	Dat *na*			1pl *mys*	
12.								Fut *er*				2sg *uŋ*		Dat *g* (+*ök*)	Dat *ne*			1pl *mis*	
13.								Fut *yr, -r*				2sg *öŋ*		Dat *x* (+*ox*)	Dat *γ* (+*ox*)			1pl *bys*	
14.								Fut.Neg *bas*				2sg *eŋ*		Dat *k* (+*ök*)	Dat *g* (+*ök*)			1pl *bis*	
15.								Fut.Neg *bes*				2sg *iŋ*		Acc *ny*	Dat *x* (+*ox*)			1pl *pys*	
16.								Fut.Neg *pas*				2sg *əŋ*		Acc *ni*	Dat *k* (+*ök*)			1pl *pis*	
17.								Fut.Neg *pes*				3sg,pl *y*		Acc *ty*	Acc *ny*			2pl *zar*	
18.								Fut.Neg *mas*				3sg,pl *a*		Acc *ti*	Acc *ni*			2pl *zer*	
19.								Fut.Neg *mes*				3sg,pl *e*		Acc *n* (+*ox,ök*)	Acc *n* (+*ox,ök*)			2pl *sar*	
20.								Hab *ǯaŋ*				3sg,pl *u*		Acc *t* (+*ox,ök*)	Acc *ty*			2pl *ser*	
21.								Hab *ǯeŋ*				3sg,pl *i*		Loc *da*	Acc *ti*			1sg *m*	
22.								Hab *čaŋ*				3sg,pl *ə*		Loc *de*	Acc *t* (+*ox,ök*)			1sg *ym*	
23.								Hab *čeŋ*				3sg,pl *ö*		Loc *ta*	Loc *da*			1sg *im*	
24.								RPast *dy*				3sg,pl *zy*		Loc *te*	Loc *de*			2sg *ŋ*	
25.								RPast *di*				3sg,pl *zi*		Loc *d* (+*ox,ök*)	Loc *nda*			2sg *yŋ*	
26.								RPast *ty*				1pl *bys*		Loc *t* (+*ox, ök*)	Loc *nde*			2sg *iŋ*	

№	0	1	2	3	4	5	6	7	8	9	10	11	12	13		14	15	16	17
27.								RPast *ti*				1pl *bis*		Abl *daŋ*	Loc *ta*			2pl *ŋar*	
28.								Past *γan*				1pl *ybys*		Abl *deŋ*	Loc *te*			2pl *ŋer*	
29.								Past *gen*				1pl *abys*		Abl *naŋ*	Loc *d* (+*ox*,*ök*)			2pl *yŋar*	
30.								Past *xan*				1pl *ubys*		Abl *neŋ*	Loc *t* (+*ox*,*ök*)			2pl *iŋer*	
31.								Past *ken*				1pl *ibis*		Abl *taŋ*	Loc *nd* (+*ox*,*ök*)			3pl *lar*	
32.								Evid *tyr*				1pl *əbis*		Abl *teŋ*	Abl *daŋ*			3pl *ler*	
33.								Evid *tir*				1pl *ebis*		All *zar*	Abl *deŋ*			3pl *nar*	
34.								Past Antipf *γalax*				1pl *öbis*		All *zer*	Abl *naŋ*			3pl *ner*	
35.								Past Antipf *gelek*				2 pl *ŋar*		All *sar*	Abl *neŋ*			3pl *tar*	
36.								Past Antipf *xalax*				2 pl *ŋer*		All *ser*	Abl *taŋ*			3pl *ter*	
37.								Past Antipf *kelek*				2 pl *yŋar*		Instr *naŋ*	Abl *teŋ*			Imp 1sg *əm*	
38.								Past Antipf *alax*				2 pl *aŋar*		Instr *neŋ**	All *zar*			Imp3sg *zyn*	
39.								Past Antipf *elek*				2 pl *uŋar*		Prol *ča*	All *zer*			Imp3sg *zin*	
40.								Conv1 *yp*				2 pl *iŋer*		Prol *če*	All *nzar*			Imp3sg *syn*	
41.								Conv1 *ip*				2 pl *əŋer*		Prol *ǯa*	All *nzer*			Imp3sg *sin*	
42.								Conv1 *p*				2 pl *eŋer*		Prol *ǯe*	All *sar*			Imp1dual *aŋ*	
43.								Conv1 *b*				2 pl *öŋer*		Prol *č* (+*ox*, *ök*)	All *ser*			Imp1dual *eŋ*	
44.								Conv1 *m*						Prol *ǯ* (+*ox*, *ök*)	Instr *naŋ*			Imp1pl *əbys*	
45.								Conv1 Neg *bən*						Delib *daŋar*	Instr *neŋ*			Imp1pl *əbis*	
46.								Conv1 Neg *pən*						Delib *deŋer*	Prol *ča*			Imp1pl Inclus *aŋar*	
47.								Conv1 Neg *mən*						Delib *naŋar*	Prol *če*			Imp1pl Inclus *eŋer*	
48.								Conv1 *abas*						Delib *neŋer*	Prol *ǯa*			Imp2pl *yŋar*	
49.								Conv1 *ebes*						Delib *taŋar*	Prol *ǯe*			Imp2pl *iŋer*	
50.								Conv2 *a*						Delib *teŋer*	Prol *nǯa*			Imp2pl *ŋar*	
51.								Conv2 *e*						Comp *taγ*	Prol *nǯe*			Imp2pl *ŋer*	
52.								PastLim *γalə*						Comp *daγ*	Prol *č* (+*ox*,*ök*)			Imp3pl *zynnar*	
53.								PastLim *gelə*						Comp *teg*	Prol *ǯ*(+*ox*,*ök*)			Imp3pl *zinner*	
54.								PastLim *xalə*						Comp *deg*	Prol *nǯ*(+*ox*,*ök*)			Imp3pl *synnar*	
55.								PastLim *kelə*							Delib *daŋar*			Imp3pl *sinner*	

№	0	1	2	3	4	5	6	7	8	9	10	11	12	13		14	15	16	17
56.								Cond *sa*							Delib *deŋer*			Prec 1sg *əmdax*	
57.								Cond *se*							Delib *naŋar*			Prec 1sg *əmdek*	
58.								Cond *za*							Delib *neŋer*			Prec 2sg *dax*	
59.								Cond *ze*							Delib *taŋar*			Prec 2sg *tax*	
60.								Conj *žyx*							Delib *teŋer*			Prec 2sg *dek*	
61.								Conj *žik*							Comp *taɣ*			Prec 2sg *tek*	
62.								Conj *čyx*							Comp *daɣ*			Prec 3sg *zyndax*	
63.								Conj *čik*							Comp *teg*			Prec 3sg *syndax*	
64.								Opt *ɣaj*							Comp *deg*			Prec3sg *zindek*	
65.								Opt *gej*							Comp *ndaɣ*			Prec 3sg *sindek*	
66.								Opt *xaj*							Comp *ndeɣ*			Prec 1dual *aŋdax*	
67.								Opt *kej*										Prec 1dual *eŋdek*	
68.								Probab *ɣadaɣ*										Prec 1pl *əbystax*	
69.								Probab *gedeg*										Prec.1pl *əbistek*	
70.								Probab *xadaɣ*										Prec.1pl Incl *aŋardax*	
71.								Probab *kedeg*										Prec.1pl Incl *eŋerdek*	
72.																		Prec.2pl *yŋardax*	
73.																		Prec.2pl *iŋerdek*	
74.																		Prec.2pl *ŋardax*	
75.																		Prec.2pl *ŋerdek*	
76.																		Prec.3pl *zynnardax*	

Explanatory notes to Table 1

The upper row is numbered by the order of the affix in Gleason's terminology. Column headings are grammatical category labels for the suffixes following Root or Stem. Cell entries contain grammatical markers that express these categories at the surface level. Each of grammatical categories may not to be expressed at the surface level, as, for example, singular number or nominative case in nouns or 3[d] person in verbs or nouns.

Abbreviations

R (Stem) – Root or Stem. Stem consists of a root and derivational affixes, included in the entries of the dictionary. The criterion for inclusion of an affix (e.g. agent affix or affix of voice or diminutive affixes) in the set of derivational morphemes is its regular presence in the dictionary's entries.

Distr – distributive, marker that indicates plurality of a subject or object of action; it may be used as a marker of the iterativeness;

Form – formative affix;

Form.Neg – negative form of formative affix;

Emph – emphatic affix;

Perf / Prosp – perfective (marker of completeness of action), prospective (marker of a state preceding an action);

Dur – durative (a marker of duration of action);

Neg – Negation;

Pres – Present tense;

Past – Past tense;

Fut.Neg – marker of negative form of the Future;

Conv – Adverbial Participle;

Conv Neg – Adverbial Participle Negative;

Hab – Habitual (Present, Past);

RPast – Recent Past;

Evid – Evidential Past;

Past Antipf – unperformed Past;

PastLim – limit in the Past (Adverbial Participle);

Cond – Conditional Mood;

Conj – Conjunctive Mood;

Opt – Optative Mood;

Probab – Hypothetical (probabilis) Mood;

Imp – Imperative Mood;

ImpInclus – Imperative inclusive (1&2 person);

Prec – Precatory Mood;

PrecInclus – Precatory inclusive (1&2 person);

Irr – subjunctive mood ("irrealis", may be combined with other tenses and moods)

Comit – comitative (marker of conformity).

Num (Sg, Pl, Dual) – number (singular, plural, dual)

Poss – possession

APos – marker of possessive attribute

Cases

Nominative (or zero case) is omitted as it has zero marker;

Gen – Genitive;

Dat – Dative;

Acc – Accusative;

Loc – Locative;

Abl – Ablative;

All – Allative;

Instr – Instrumental;

Prol – Prolative;

Delib – Deliberative (case of indirect object);

Comp – Comparative;

Attr – marker of Attribute;

Adv – marker of Circumstantial modifier.

№	Phonetic characteristics of the previous element	Plural
1.	a) Among the elements preceeding the affix the last vowel is a back one (*a, y, o, u*) & b) the previous element terminates in a vowel or in a voiced non-nasal consonant (*b, v, γ, d, ž, j, z, l, r*)	*lar*
2.	a) Among the elements preceeding the affix the last vowel is a front one (*e, ə, i, ö, ü*) & b) the previous element terminates in a vowel or in a voiced non-nasal consonant (*b, v, g, d, ž, j, z, l, r*)	*ler*
3.	a) Among the elements preceeding the affix the last vowel is a back one (*a, y, o, u*) & b) the previous element terminates in a vowel or in a non-voiced consonant (*p, f, x, t, š, s, c, č*) or in devocalized voiced consonant (*b, v, d, ž, z*) in Russian loan-words (*zavod-tar*).	*tar*
4.	a) Among the elements preceeding the affix the last vowel is a front one (*e, ə, i, ö, ü*) & b) the previous element terminates in a vowel or in a non-voiced consonant (*p, f, k, t, š, s, c, č*) or in devocalized voiced consonant (*b, v, d, ž, z*) in Russian loan-words.	*ter*
5.	a) Among the elements preceeding the affix the last vowel is a back one (*a, y, o, u*) & b) the previous element terminates in a nasal consonant (*m, n, ŋ*)	*nar*
6.	a) Among the elements preceeding the affix the last vowel is a front one (*e, ə, i, ö, ü*) & b) the previous element terminates in a nasal consonant (*m, n, ŋ*)	*ner*

Table 2: Rules for choice of affixes of Plural

8. Conclusion

In this paper we describe the preliminary linguistic work for creating automatic morphological annotation for the Khakass written corpus. Main components of this work are: 1) the database of Khakass stems constructed with the StarLing system, 2) the computational model of the Khakass wordform based on the combinatorial principles known as morpheme order (Gleason (1955) and 3) the set of phonetic rules that constrain the choice of allomorphs within the wordform. We also presented the list of Khakass inflectional affixes with their allomorphs.

The present model of a wordform and the set of rules are intended to be evaluated on a large and diverse corpus of Khakass, and subsequently modified and supplemented.

9. References[3]

Baskakov 1953 – Баскаков Н.А. Очерк «Хакасский язык» // Хакасско-русский словарь / Под ред. Н.А.Баскакова. М., 1953. {Baskakov N.A. Ocherk «Hakasskij jazyk» // Hakassko-russkij slovar' / Pod red. N.A.Baskakova. M., 1953}

Çağrı Çöltekin (2010). A Freely Available Morphological Analyzer for Turkish / http://www.let.rug.nl/coltekin/papers/coltekin-lrec2010.pdf; http://www.let.rug.nl/coltekin/trmorph/

[3] At present we don't know any linguistic descriptions of Khakass in any languages other than Russian.

Gleason, H. (1955). Introduction to descriptive linguistics. New York, 1955: Holt, Rinehart and Winston.

Khakass Grammar 1975 – Грамматика хакасского языка / Под ред. Н.А.Баскакова. М., 1975. {Grammatika hakasskogo jazyka / Pod red. N.A.Baskakova. M., 1975}

Krylov 2008 – Крылов С.А. Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии // Orientalia et classica. Выпуск XIX. Аспекты компаративистики. 3. М., РГГУ, 2008, с. 649–668. {Krylov S.A. Strategii primenenija integrirovannoj informacionnoj sredy StarLing v korpusnoj lingvistike i v komp'juternoj leksikografii // Orientalia et classica. Vypusk XIX. Aspekty komparativistiki. 3. M., RGGU, 2008, s. 649–668.}

Krylov 2011 – Крылов С.А. Использование системы StarLing при создании морфологически аннотированного корпуса современного монгольского языка. (printed as manuscript). {Krylov S.A. Ispol'zovanie sistemy StarLing pri sozdanii morfologicheski annotirovannogo korpusa sovremennogo mongol'skogo jazyka.}

Large Khakass-Russian Dictionary (2006) – Большой хакасско-русский словарь / Под ред. О.В.Субраковой. Новосибирск, «Наука». {Bol'shoj hakassko-russkij slovar' / Pod red. O.V.Subrakovoj. Novosibirsk, «Nauka».}

Lyashevskaya et al. 2005 – Ляшевская О.Н., Плунгян В.А., Сичинава Д.В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. – М., 2005. 111–135. {Ljashevskaja O.N., Plungjan V.A., Sichinava D.V. O morfologicheskom standarte Nacional'nogo korpusa russkogo jazyka // Nacional'nyj korpus russkogo jazyka: 2003–2005. Rezul'taty i perspektivy. – M., 2005. 111–135.}

Mal'tseva 2004 – Мальцева В.С. Структура глагольной словоформы в сагайском диалекте хакасского языка (говор с. Казановка) / Дипломная работа (graduation work; printed as manuscript). М., 2004. {Mal'tseva V.S. Struktura glagol'noj slovoformy v sagajskom dialekte hakasskogo jazyka (govor s. Kazanovka). M., 2004.}

Revzin, Yuldasheva 1969 – Ревзин И.И., Юлдашева Г.Д. Грамматика порядков и ее использование.// Вопросы языкознания, №1. с. 42–56. {Revzin I.I., Juldasheva G.D. Grammatika porjadkov i ee ispol'zovanie.// Voprosy jazykoznanija, №1. s. 42–56.}

Сиразитдинов З.А. Алгоритмическая грамматика словоизменения башкирского языка / http://mfbl.ru/bashdb/algram/algram.htm {Sirazitdinov Z.A. Algoritmicheskaja grammatika slovoizmenenija bashkirskogo jazyka / http://mfbl.ru/bashdb/algram/algram.htm}

Volodin, Khrakovsky 1975 – Володин А.П., Храковский В.С. Типология морфологических классификаций глагола (на материале агглютинативных языков) // Типология грамматических категорий: Мещаниновские чтения. М.: Наука, 1975. {Volodin A.P., Hrakovskij V.S. Tipologija morfologicheskih klassifikacij glagola (na materiale aggljutinativnyh jazykov) // Tipologija grammaticheskih kategorij: Mewaninovskie chtenija. M.: Nauka, 1975.}

Online Resources:

[1] http://www.eki.ee/books/redbook/Khakass.shtm
[2] http://ru.wikipedia.org/wiki/Тюркские_языки
[3] http://corpling-ran.ru/n3.html
[4] http://www.sanasoft.kz/a/node/52
[5] http://www.azerispellcheck.com/AzeriSpeller.asp
[6] http://mfbl.ru/ bashdb/algram/algram.html
[7] http://starling.rinet.ru/program

Annotating Cognates and Etymological Origin in Turkic Languages

Benjamin S. Mericli*, Michael Bloodgood†

*University of Maryland, College Park, MD bmericli@umd.edu
†University of Maryland, College Park, MD meb@umd.edu

Abstract

Turkic languages exhibit extensive and diverse etymological relationships among lexical items. These relationships make the Turkic languages promising for exploring automated translation lexicon induction by leveraging cognate and other etymological information. However, due to the extent and diversity of the types of relationships between words, it is not clear how to annotate such information. In this paper, we present a methodology for annotating cognates and etymological origin in Turkic languages. Our method strives to balance the amount of research effort the annotator expends with the utility of the annotations for supporting research on improving automated translation lexicon induction.

1. Introduction

Automated translation lexicon induction has been investigated in the literature and shown to be feasible for various language families and subgroups, such as the Romance languages and the Slavic languages (Mann and Yarowsky, 2001; Schafer and Yarowsky, 2002). Although there have been some studies investigating using Swadesh lists of words to identify Turkic language groups and loanword candidates (van der Ark et al., 2007), we are not aware of any work yet on automated translation lexicon induction for the Turkic languages.

However, the Turkic languages are well suited to exploring such technology since they exhibit many diverse lexical relationships both within family and to languages outside of the family through loanwords. For the Turkic languages, it is prudent to leverage both cognate information and other etymological information when automating translation lexicon induction. However, we are not aware of any corpora for the Turkic languages that have been annotated for this information in a suitable way to support automatic translation lexicon induction. Moreover, performing the annotation is not straightforward because of the range of relationships that exist. In this paper, we lay out a methodology for performing this annotation that is intended to balance the amount of effort expended by the annotators with the utility of the annotations for supporting computational linguistics research.

2. Main Annotation System

We obtained the dictionary of the Turkic languages (Öztopçu et al., 1996). One section of this dictionary contains 1996 English glosses and for each English gloss a corresponding translation in the following eight Turkic languages: Azerbaijani, Kazakh, Kyrgyz, Tatar, Turkish, Turkmen, Uyghur, and Uzbek. Table 1 shows an example for the English gloss 'alive.' When a language has an official Latin script, that script is used. Otherwise, the dictionary's transliteration is shown in parentheses. Our annotation system is to annotate each Turkic word with a two-character code. The first character will be a number indicating which words are cognate with each other and the second character will indicate etymological information. Subsection 2.1. discusses how to define and annotate cognates and subsection 2.2. discusses how to define and annotate etymological information.

2.1. Cognates

According to the Oxford English Dictionary Online[1] accessed on February 2, 2012, 'cognate' is defined as: "...Of words: Coming naturally from the same root, or representing the same original word, with differences due to subsequent separate phonetic development; thus, English *five*, Latin *quinque*, Greek πενταε, are cognate words, representing a primitive **penke*." As this definition shows, shared genetic origin is key to the notion of cognateness. A word is only considered cognate with another if both words proceed from the same ancestor. Nonetheless, in line with the conventions of previous research in computational linguistics, we set a broader definition. We use the word 'cognate' to denote, as in (Kondrak, 2001): "...words in different languages that are similar in form and meaning, without making a distinction between borrowed and genetically related words; for example, English 'sprint' and the Japanese borrowing 'supurinto' are considered cognate, even though these two languages are unrelated." These broader criteria are motivated by the ways scientists develop and use cognate identification algorithms in natural language processing (NLP) systems. For cross-lingual applications, the advantage of such technology is the ability to identify words for which similarity in meaning can be accurately inferred from similarity in form; it does not matter if the similarity in form is from strict genetic relationship or later borrowing.

However, not every pair of apparently similar words will be annotated as cognate. For them to be considered cognates, the differences in form between them must meet a threshold of consistency within the data. We will explain the definitions and rules for the annotators to follow in order to establish such a threshold.

First, we elaborate on how our notion of cognate differs from that of strict genetic relation. At a high level, there are two cases to consider: A) where the words involved are native Turkic words, and B) where the words involved are

[1] http://www.oed.com/view/Entry/35870?redirectedFrom=cognate

shared loanwords from non-Turkic languages. Within case A, there are two cases to consider: (A1) genetic cognates; and (A2) intra-family loans. Table 2 shows an example of case A1. This example shows the English gloss 'one' for all eight Turkic languages, descended from the same postulated form, **bir*, in Proto-Turkic (Róna-Tas, 2006). Case A1 is the strict definition of 'cognate,' and these are to be annotated as cognate.

Case A2 is for intra-family loans, i.e., a word of ultimately Turkic origin borrowed by one Turkic language from another Turkic language. These cases, contrary to the strict definition, are to be marked as cognate in our system. An example is the modern Turkish neologism *almaş* 'alternation, permutation', incorporated from the Kyrgyz (*almaş*) 'change' (Türk Dil Kurumu, 1942). While rare, it is used today in Turkish scholarly literature to describe concepts in areas such as mathematics and botany. Processing genetic cognates (case A1) and intra-family loans (case A2) differently would have little impact on the success of a cross-dictionary lookup system. In fact, accounting for the difference might limit the efficacy of such a system. Also, the time depth of intra-Turkic borrowings may be centuries or mere decades. The more distant the borrowing the more difficult it will be for annotators to distinguish between cases A1 and A2. Hence, instances of case A2 are to be annotated as cognate in our system.[2]

Case B is for situations of shared loanwords, where the source of the words is ultimately non-Turkic. There are three subcases: (B1) loanwords borrowed from the same non-Turkic language; (B2) loanwords borrowed from different non-Turkic languages, but of the same ultimate origin; and (B3) loanwords of non-Turkic origin borrowed via another Turkic language.

Table 3 shows an example of case B1, the word 'book,' borrowed from Arabic in all eight Turkic languages. Table 4 shows an example of case B2, the word 'ballet,' borrowed from Russian in all cases except Turkish, where it was borrowed directly from the French. Table 5 shows an example of case B3: the word 'benefit' in Kyrgyz was borrowed most likely through Uzbek or Chaghatay (Kirchner, 2006), but the Uzbek word was borrowed from Persian, and ultimately from Arabic. It is difficult and time-consuming for annotators to make these fine-grained distinctions. And again, for computational processing, such distinctions are not expected to be helpful. Hence, all of cases B1, B2, and B3 are to be annotated as cognate in our system.

Recall that all our annotations are two-character codes; the first character is a number from one to eight indicating what words are cognate with each other. Table 6 shows the first character of the annotations for the example from Table 1. The words marked with 1 are cognate with each other and the words marked 2 are cognate with each other.

[2]For similar reasons, false cognates may be annotated as cognate if the annotator does not have readily available knowledge indicating that they are false cognates. Although this is a potential limitation of our system, it is not clear how to distinguish false cognates from true cognates without significant additional annotation expense.

2.2. Etymology

The second character in a word's annotation indicates a conjecture about etymological origin, e.g., T for Turkic. The decision to annotate word origin is motivated by its value for facilitating the development of technology for cross-language lookup of unknown forms. We therefore take a practical approach, balancing the value of the annotation for this purpose with the amount of effort required to perform the annotation. We have created the following code for annotating etymology:

T Turkic origin. This includes compound forms and affixed forms whose constituents are all Turkic. For example, the Turkmen for 'manager', *ýolbaşçy*, is marked T because its compound base, *ýol* with *baş*, and affix *-çy* are all Turkic in origin.

A Arabic origin, to include words borrowed indirectly through another language such as Persian. For example, the word in every Turkic language for 'book' is marked A for all eight Turkic languages. Because variations on the Arabic form /kita:b/ exist in every Turkic language, in Persian, and in other languages of the Islamic world, it is difficult to tease out the word's trajectory into a language such as Kyrgyz. The burden of researching these fine distinctions is not placed on the annotator, as explained below.

P Persian origin, not including Arabic words possibly borrowed through Persian. An example is the word for 'color' in many Turkic languages, from the Persian /ræng/.

R borrowed from Russian, including words that are ultimately of French origin.

F French origin, not including ultimately French words borrowed from Russian. Direct French loans occur almost exclusively in Turkish. An example is the word for 'station' in Turkish, *istasyon*.

E English origin. For example the word for 'basketball' in every language.

I Italian origin. Usually of importance only to specific domains in Turkish.

G Greek origin. For example, the word in Azerbaijani, Turkish, Turkmen, Uyghur, and Uzbek for 'box' comes from the Greek κουτι.

C Chinese origin, usually Mandarin and usually of importance only to Uyghur. An example is the word for 'mushroom' in Uyghur, (*mogu*).

Q unknown or inconclusive origin.

The careful reader will have noticed that there is an inconsistency in that words of ultimately Arabic origin borrowed through Persian are marked as A, but words of ultimately French origin borrowed through Russian are marked as R. There are two reasons for this. The first is annotator efficiency. Making the judgment that a word is ultimately of Arabic origin is much easier than having to figure out

whether it was borrowed from Arabic or indirectly from Persian. For the Russian/French situation, the distinction is much easier to make. To begin with, the Russian loanwords occur almost exclusively in former USSR languages and the French loanwords occur almost exclusively in Turkish. Also, the orthography often gives clear cues for making this distinction, as Russian loanwords consistently retain characteristically Russian letters.

2.2.1. Multi-Language Exceptions

We also define other codes that categorize certain complex words that do not fall into any of the categories described in subsection 2.2.. Other etymological annotation studies, such as the Loanword Typology project and its World Loanword Database (Haspelmath and Tadmor, 2009), have instructed linguists to pass over such complex words and optionally flag them as "contains a borrowed base," etc. Our annotation system requires that these words, which are very common in Turkic languages, be annotated according to more fine grained categories.

The following are our multi-language exception codes:

X Compound words where the constituents are from different origins. For example, the Tatar word for 'truck', (*yök mashinası*), is to be marked X since it contains Russian-origin (*mashina*), 'machine, vehicle' in compound with Tatar (*yök*), 'baggage,cargo.' In contrast, the Turkish compound word for thunder, *gök gürlemesi*, will be marked T because all of its constituents are Turkish.

V A verb formed by combining a non-Turkic base with a Turkic auxiliary verb or denominal affix. For example, the verb 'to repeat' in Azerbaijani, Tatar, and Turkish, because it consists of a noun borrowed from the Arabic /takra:r/ plus a Turkic auxiliary verb *et-* or *it-*.

N A nominal consisting of a non-Turkic base bearing one or more Turkic affixes, in cases where removing the affixes results in a form that can plausibly be found elsewhere in the data or in a loan language dictionary. For example, the Kazakh word for 'baker,' (*nawbayshı*), is composed of a Persian-origin base, from /na:nva:/, 'baker', and a suffix that indicates a person associated with a profession, (*-shı*). The Turkmen word for 'baker,' (*çörekçi*), on the other hand, will be marked T, because both its base (*çörek*) and affix (*-çi*) are Turkic.

Table 7 shows an example of an entry that has been fully annotated for both cognates and etymology.

3. Inter-Annotator Agreement

We pilot-tested our annotation system with two annotators on 400 etymology annotations.[3] Both annotators have studied linguistics. Also, both are native English speakers with experience studying or speaking multiple Turkic languages, Persian, and Arabic. Training consisted of studying the authors' annotation manual and asking any follow-up questions. Both annotators made approximately 240 annotations per hour.

[3]Table 8 has 392 entries because the annotators claimed eight entries had multiple translations for the same English gloss.

Table 8 shows the contingency matrix for annotating the 400 entries.[4] From Table 8 it is immediate that agreement is substantial, and when there is disagreement it is largely for the difficult cases of inconclusive origin and the multi-language exceptions: Q, X, V, and N. We measured inter-annotator agreement using Cohen's Kappa (Cohen, 1960) and found Kappa = 0.5927 (95% CI = 0.5192 to 0.6662). If we restrict attention to only the instances where neither of the annotators marked an inconclusive origin or multi-language exception, then Kappa is 0.9216, generally considered high agreement. This shows that our annotation system is feasible for use and also shows that to improve the system we might focus efforts on finding ways to increase agreement on the annotation of the exceptional cases (Q, X, V, and N).

	T	A	P	R	F	Q	X	V	N
T	160	8	2	0	0	3	10	6	1
A	0	56	2	6	0	1	0	1	0
P	0	0	31	0	0	0	1	0	0
R	0	0	0	32	1	0	0	0	0
F	0	0	0	0	5	0	0	0	0
Q	12	5	0	2	0	0	2	3	0
X	2	0	1	5	0	0	17	8	0
V	0	1	0	0	0	0	0	0	0
N	0	0	1	0	0	0	6	0	1

Table 8: Table of Counts for two annotators' etymological conjectures on 392 words. Annotator 1's conjectures follow the horizontal axis, and annotator 2's the vertical.

4. Conclusions and Future Work

The Turkic languages are a promising candidate family of languages to benefit from automated translation lexicon induction. A necessary step in that direction is the creation of annotated data for cognates and etymology. However, this annotation is not straightforward, as the Turkic languages exhibit extensive and diverse etymological relationships among words. Some distinctions are difficult for annotators to make and some are easier. Also, some distinctions are expected to be more useful than others for automating cross-lingual applications among the Turkic languages. We presented an annotation methodology that balances the research effort required of the annotator with the expected value of the annotations. We surveyed and explained the wide range of the most important relationships observed in the Turkic languages and how to annotate them. When we finish the annotations, we would like to make the annotated data available as long as it is legal under copyright laws for us to do so. Finally, we hope that our annotation system and the associated discussion can be useful for other teams that are annotating Turkic resources, and perhaps parts of it can be useful for annotating resources for other language families as well.

[4]We left out columns for English, Greek, Italian, and Chinese, which were not relevant for the 50 entries (according to unanimous agreement of our annotators).

5. References

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Martin Haspelmath and Uri Tadmor. 2009. The Loanword Typology project and the World Loanword Database. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 1–34, Berlin. Walter de Gruyter.

Mark Kirchner. 2006. Kirghiz. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 344–356, New York. Routledge.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kurtuluş Öztopçu, Zhoumagaly Abuov, Nasir Kambarov, and Youssef Azemoun. 1996. *Dictionary of the Turkic Languages*. Routledge, New York.

András Róna-Tas. 2006. The reconstruction of Proto-Turkic and the genetic question. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 67–80, New York. Routledge.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Türk Dil Kurumu. 1942. *Felsefe ve Gramer Terimleri*. Cumhuriyet Basımevi, Istanbul.

René van der Ark, Philippe Mennecier, John Nerbonne, and Franz Manni. 2007. Preliminary identification of language groups and loan words in Central Asia. In *Proceedings of the RANLP Workshop on Computational Phonology*, pages 12–20, Borovetz, Bulgaria.

Azerbaijani	Kazakh	Kyrgyz	Tatar	Turkish	Turkmen	Uyghur	Uzbek
canlı	(tiri)	(türüü)	(janlı)	canlı	diri	(tirik)	tirik

Table 1: Example entry from the eight-way dictionary for the English gloss ‘alive.’

Azerbaijani	Kazakh	Kyrgyz	Tatar	Turkish	Turkmen	Uyghur	Uzbek
bir	(bir)	(bir)	(ber)	bir	bir	(bir)	bir

Table 2: Example of case A1: genetic cognates. The English gloss is ‘one.’

Azerbaijani	Kazakh	Kyrgyz	Tatar	Turkish	Turkmen	Uyghur	Uzbek
kitab	(kitap)	(kitep)	(kitap)	kitap	kitap	(kitab)	kitob

Table 3: Example of case B1: loanwords borrowed from the same non-Turkic language. The English gloss is ‘book.’

Azerbaijani	Kazakh	Kyrgyz	Tatar	Turkish	Turkmen	Uyghur	Uzbek
balet	(balet)	(balet)	(balet)	bale	balet	(balet)	balet

Table 4: Example of case B2: loanwords borrowed from different non-Turkic languages, but of the same ultimate origin. The English gloss is ‘ballet.’

Azerbaijani	Kazakh	Kyrgyz	Tatar	Turkish	Turkmen	Uyghur	Uzbek
fayda	(payda)	(payda)	(fayda)	fayda	peýda	(payda)	foyda

Table 5: Example of case B3: loanwords of non-Turkic origin borrowed via another Turkic language. The English gloss is ‘benefit.’

Azerbaijani	Kazakh	Kyrgyz	Tatar	Turkish	Turkmen	Uyghur	Uzbek
canlı	(tiri)	(türüü)	(janlı)	canlı	diri	(tirik)	tirik
1	2	2	1	1	2	2	2

Table 6: Example with cognates annotated.

Azerbaijani	Kazakh	Kyrgyz	Tatar	Turkish	Turkmen	Uyghur	Uzbek
stul	(orındıq)	(orunduk)	(urındık)	sandalye	stul	(orunduq)	kursi
1R	2T	2T	2T	3A	1R	2T	4A

Table 7: Example with complete annotation both for cognates and etymology. The English gloss here is ‘chair.’

Semi-supervised morpheme segmentation without morphological analysis

Özkan Kılıç, Cem Bozşahin
Department of Cognitive Science
Informatics Institute, Middle East Technical University
Ankara, Turkey
E-mail: okilic@ii.metu.edu.tr, bozsahin@metu.edu.tr

Abstract

The premise of unsupervised statistical learning methods lies in a cognitively very plausible assumption that learning starts with an unlabeled dataset. Unfortunately such datasets do not offer scalable performance without some semi-supervision. We use 0.25% of METU-Turkish Corpus for manual segmentation to extract the set of morphemes (and morphs) in its 2 million word database without morphological analysis. Unsupervised segmentations suffer from problems such as oversegmentation of roots and erroneous segmentation of affixes. Our supervision phase first collects information about average root length from a small fragment of the database (5,010 words), then it suggests adjustments to structure learned without supervision, before and after a statistically approximated root, in an HMM+Viterbi unsupervised model of n-grams. The baseline of .59 f-measure goes up to .79 with just these two adjustments. Our data is publicly available, and we suggest some avenues for further research.

1. Introduction

Morpheme segmentation is the process of revealing the morphs or morphemes in a word. It can be conceived in two ways: (i) providing a sequence of morphosyntactic tags associated with the entire word, (ii) dividing the word into its morphs, with some morphic or morphemic tagging to go along with the substring covering the morph. The most common way for both tasks so far has been through morphological analysis. We describe in this work a way to approach the problem in (ii) without analysis or morphological parsing, which we tested on Turkish. As far as we know, this has been done for the first time. It uses a common Turkish language resource in two ways for the evaluation of segmenting the words into their morphemes. The first phase, semi-supervision, also yielded a gold standard of manual segmentation without labels, which we make publicly available. Although the resource is small in size (10,582 morphemes of 5,010 words), its contribution to the task is very significant, and it provides a common base for comparison in the future because the words are drawn from a well-known resource.

Rule-based morphological analyzers employ finite-state approaches with a previously compiled lexicon of morphemes. They use a set of rules for language-specific morphotactics and morpho-phonological constraints. They have been applied to concatenating languages (Koskenniemi, 1983; Hankamer, 1986; Oflazer, 1994; Çöltekin, 2010) and nonlinear templatic languages (Kiraz, 2002; Cohen-Sygal et al., 2003). Such methods are language-specific, and require their lexicons and rule sets to be updated. Statistical approaches to morpheme segmentation depend on the training of hypothetical models, which requires excessive amounts of data, from few hundred thousand to millions of words. There are well-known methods, namely supervised methods (Hajic & Hladka, 1997; Hakkani-Tür et al., 2002), unsupervised methods (Baroni et al., 2002; Creutz & Lagus, 2005; Goldwater, 2007; Yatbaz & Yuret, 2009; Yatbaz & Yuret, 2010), and semi-supervised ones (Yarowsky & Wicentowski, 2002; Kohonen et al., 2010). In these methods the training data are labeled, unlabeled, or partially labeled respectively.

Turkish is an agglutinating language with a complex morphology. Precise modeling of its morphemes using statistical methods requires large amount of data. The available resources are the METU-Turkish Corpus (Say et al., 2002) and similar academic corpora (Sak et al., 2011). This study describes an application of the Hidden Markov Model (HMM), an unsupervised method, to two million words of METU-Turkish Corpus in the first stage. The morphological tags of the corpus are ignored for unsupervised learning, and no morpheme segmentation and syntactic annotation are employed. The n-grams that form the basis for HMM are defined as states with respect to their orthographic lengths; and possible collections of orthographic representations for each n-gram are defined as emissions.

In the second stage, the model is trained on the corpus of orthographic representations of 5,010 words selected from the corpus, to calculate the initial transition and emission probabilities. These words are manually segmented by us, giving the ratio of 2 million words to 5,010 in unsupervised and supervised training. Viterbi algorithm was employed to find the most probable segmentation of a given word.

The Viterbi algorithm might suffer from the local maxima problem of the HMM. The local maxima may result from an ambiguous orthographic representation cluster which looks like a morpheme (or more precisely, a morph). It is mainly because of the tension between contrast and efficiency. Optimizing both elements gives rise to ambiguities in the collocations. For example, most of the derivational affixes and some of the inflectional affixes are frequently polysemous. For example, the suffix *-lAr* in Turkish functions as both Plu and 3P.Plu. The stem *anla* 'understand' terminates with a segment which is homographic with the inflectional suffix *-la*

(Instrumental). Similarly, the stem *ak* 'white' and the suffix *-ak* (a derivational suffix) are homographs. (Since we work on orthographic representations, we lack the phonological information such as stress to disambiguate them as homophones). As a result of such ambiguities, false segmentations such as dividing *-lar* (Plu) into *-la* (Ins) and *-r* (Aorist), and oversegmentations, e.g. dividing *kiler* 'pantry' into *-ki* (Relative) and *-ler* (Plu) do occur in the unsupervised method. Our manual segmentation of the small fragment of the database is intended to see how we can cope with these problems without attempting a morphological analysis of the test data. The method, improvements and our findings are described in the subsequent sections.

2. Method

Our HMM is a statistical model which is used to evaluate the probability of a sequence of morphemes. The model uses the Markov chain property:

• $P(s_{i,k} \mid s_{i,1}, s_{i,2}, \ldots, s_{i,k-1}) = P(s_{i,k} \mid s_{i,k-1})$

Thus the probability of next state depends only on the previous state. This seems to be a simple base to start experimenting with learning concatenative morphology. In Turkish morphotactics, the continuation of a morpheme is determined by the most recent suffix attached to a stem. For example, the suffix *-ki* can only be attached to words with either Gen or Loc, to form pronominal expressions, including inherently locative-temporal nouns such as *sabah* (morning) and *akşam* (night): *ev-de-ki* (house-Loc-ki), *ev-in-ki* (house-Gen-ki), **ev-e-ki* (house-Dat-ki), *sabah-ki* and such (Bozsahin, 2002).

In the current study, the set of states are n-grams starting from unigrams up to the longest word, and the transition probabilities are the likelihoods of possible n-gram collocations. To make the calculations easier, the 'Start' and 'End' states are inserted for each word. The emission probability of an n-gram of length-*x* is evaluated through the possible orthographic representations of length-*x* in the corpus. The Viterbi algorithm finds the optimal segmentation through the probabilities of the possible paths of the states and their emissions.

2.1 Data preparation

A subset of the METU-Sabancı Turkish Treebank (Atalay et al., 2003; Oflazer et al., 2003) is manually segmented (5,010 words). The Treebank itself consists of 7,262 annotated sentences with 43,571 words from the corpus. Both derivational and inflectional affixes are segmented. The allomorphs, such as the plural suffixes *-lar* and *-ler,* or derivational suffixes *-lik* and *-liğ,* are treated as different morphs.

In the manually segmented set, the segments with respect to their orthographic lengths correspond to n-grams in the HMM. The orthographic distributions of the n-grams lead to the emissions probabilities in the HMM. For example, the segments *-lar* and *-ın* correspond respectively to the trigram (N3) and the bigram (N2) in the HMM, and a collocation *-lar-ın* is used in estimating the transition probability from N3 to N2. In a similar manner, the orthographic representations in the manually segmented set "*-ler*, *-lar, -lik…*" and "*-in, -ın, -ün …"* are possible emissions of N3 and N2.

The statistics from the manual segmentation are used to improve the model by attempting to reduce the number of false segmentations and oversegmentations.

3. Findings

3.1 Results from the HMM

We start with the naive method of exhaustive generation of possible n-grams from the Turkish alphabet, which consists of 29 letters. No phonological filtering is applied to the n-grams before evaluating their frequencies.

The frequencies speak for themselves. For example, the most frequent n-grams in this group are inflectional morphemes, as well as some connectives and frequent function words, such as *-lar* (Plu), *ve* 'and' and *bir* 'a/one'. The least frequent n-grams are usually rare stems and nonce words, such as *ihya* 'enliven', *zzzt* and *ğaü*. Table 1 provides a summary.

	Unigram	Bigram	Trigram	Tetragram
Total Types	29	779	8,948	35,628
Total Tokens ~	20 million	7.5 million	6.5 million	5 million

Table 1: Total Numbers of Observed Types and Tokens of N-grams (N ≤ 4).

The most frequent 10 tokens and their percentages from about 2 million words in the corpus are given in Table 2.

Order	Unigram	Bigram	Trigram	Tetragram
1	*a*	*ar*	*lar*	*ları*
2	*e*	*la*	*ler*	*leri*
3	*n*	*an*	*eri*	*erin*
4	*r*	*er*	*arı*	*ında*
5	*i*	*le*	*bir*	*arın*
6	*l*	*in*	*ara*	*inde*
7	*k*	*de*	*nda*	*iyor*
8	*d*	*en*	*yor*	*nlar*
9	*ı*	*ın*	*ini*	*anal*
10	*m*	*da*	*ını*	*asın*
Percent in Total Tokens	53%	16.5%	6.8%	4.1%

Table 2: Most Frequent N-grams (N ≤ 4).

Figure 1 shows a very simple trellis diagram indicating the possible state transitions for the word *kedim* 'my cat', which also corresponds to possible segmentations. The emission probabilities of each n-gram and the transition probabilities among corresponding n-grams are given in Table 3 and Table 4 respectively.

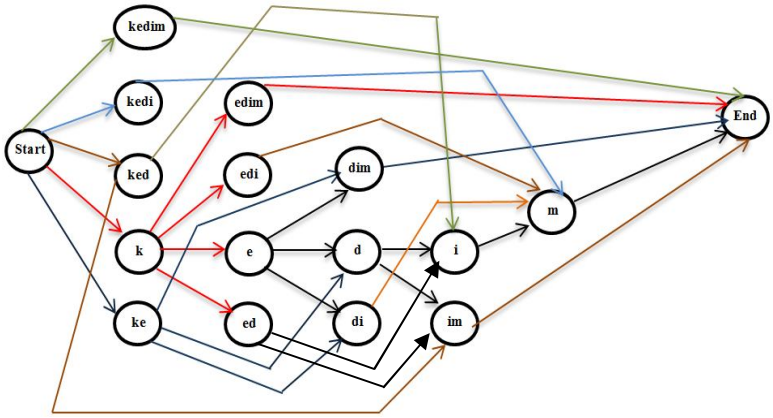

Figure 1: Trellis diagram for *kedim* 'my cat'

Output	N5	N4	N3	N2	N1	Start	End
kedim	3.81E-06						
kedi		4.73E-05					
edim		2.31E-04					
ked			1.04E-04				
edi			3.30E-03				
dim			5.43E-04				
ke				4.40E-03			
ed				4.99E-03			
di				9.22E-03			
im				4.93E-03			
k					5.23E-02		
e					7.62E-02		
d					5.08E-02		
i					7.06E-02		
m					4.12E-02		
ε						1.00E00	1.00E00

Table 3: Emission probabilities of the n-grams in the trellis diagram (Empty cells are zero. ε is empty string)

	Start	*kedim*	*kedi*	*edim*	*ked*	*edi*	*dim*	*ke*	*ed*	*di*	*im*	*k*	*e*	*d*	*i*	*m*	End
Start		1.82E-05	8.66E-03		1.87E-03			1.34E-01				4.02E-01					
kedim																	2.50E-01
kedi																1.67E-01	
edim																	2.10E-02
ked											6.02E-03				3.60E-01		
edi																5.54E-02	
dim																	2.65E-01
ke							1.19E-04			7.10E-03				1.97E-02			
ed											3.05E-02				5.51E-01		
di																4.90E-02	
im																	4.26E-01
k				1.03E-05		6.12E-04			1.70E-03				8.62E-02				
e							2.05E-03			3.70E-02				6.73E-02			
d											9.12E-03				1.86E-0		
i																7.14E-02	
m																	1.75E-01
End																	

Table 4: Transition probabilities of the n-grams in the trellis diagram (Empty cells are zero)

The Viterbi algorithm chooses the path as (Start, N4, N1, End) emitting (ε, *kedi*, *m*, ε), in which ε denotes the empty string. This is the correct sequence of morphs in the word. The second most probable path, which is slightly closer to the first path in score, is (Start, N2, N2, N1, End) emitting (ε, *ke*, *di*, *m*, ε), because of the high number of occurrences of the past tense suffix *-di* in the corpus. This is a wrong segmentation. A corpus with significantly more verbs than nouns would make the second path the winning alternative. We tried to avoid overfitting by using a representative distribution of nouns and verbs (1,414 verbs, 3,596 nouns, adjectives, adverbs and connectives). The precision, recall and f-measure values of the unsupervised method are .51, .72 and .59 respectively, which are, of course, not satisfactory.

3.2 Enhancing the Model by Manual Segmentation

The average root length of the subset we used from the Treebank is 4.09. Güngör (2003) reports the average root length to be 4.02 for Turkish. There are 150 derivational and 214 inflectional morphemes in our subset. This is the gold standard for our subset. The inflectional suffixes are very frequent. Derivational suffixes are not nearly frequent. For example, in the segmentation of the first 100 words, 59 new morphemes are discovered, of which only 6 are derivational.

To understand the cause of oversegmentations of roots by the HMM, the statistics of distinct roots whose endings are identical to morphemes in our gold standard have been evaluated from the Treebank, as shown in Table 5. For example, the most frequent root termination has the ending *-n* (10.82%).

Root Ending Segment	Percent in the Treebank
n	10.82%
k	10.13%
t	9.59%
a	9.56%
e	8.25%
r	7.69%
i	6.02%
et	4.71%
m	4.60%
an	3.86%
ş	3.18%
ı	3.04%
ol	2.41%
la	2.32%
u	2.32%
er	2.14%
le	2.00%

Table 5: Percentages of some root endings with morpheme-like segments

We incorporate this edge statistic to our experiments as follows: if the sum of the indices of visited states (a measure of length) is close to the calculated average root length 4.09, and if in the current state a symbol identical to one of our morpheme endings *x* from Table 5 is observed, then the state's transition probability is multiplied by (1- percentage-of-*x*), which gives the probability of *x* not being an edge of the roots from the Treebank. For example, if a unigram is in the 4th orthographic position of a word and it emits *-n*, then its transition probability is multiplied by (1 - 0.1082). This is a simple way to check the effect of the edge statistic on oversegmentation of roots, because it forces the Viterbi algorithm to favor likely endings of roots and morphemes.

Next we tackle the false segmentation problem of morphemes. The statistics from the segmented subset are used for this purpose to look at structure past the average root length. For example, *-lArI* (3Plu.Poss) and *-lar-I* (Plu-Acc) are identical orthographically, hence they are prone to false segmentation. Manual segmentations show that there are 190 occurrences of the latter one, of which 59% have at least one more segment before the word boundary. On the other hand, 3Plu.Poss occurs in 40 words of which 30% are in word boundaries.

The statistics of such problematic cases were part of our experiments. Their (1- 'edge probabilities') are multiplied with the transition probabilities of the HMM considering the locations and emission types of the states. For example, if *-lArI* has the transition probability .085, and *-lAr* .075, and if 70% of *-lArI* are not at the word boundary compared to 59% for *-lAr*, determined from supervision, the numbers (1-.7)x.085 and (1-.59)x.075 would be the contenders. By doing so, the Viterbi algorithm is partially directed to a path starting with a 3-gram (Plu) instead of a 4-gram (3Plu.Poss) for *-lArI-* representations occurring before the word boundaries.

4. Results and Conclusion

The working principles in our two experiments are to disfavor oversegmentations of roots and false segmentations of affixes by incorporating the collocations of root endings and morpheme starts. Employing this much semi-supervision from a very small fragment (0.25%) of the database successfully increased the measures to (.72, .87, .79) (precision, recall, f-measure), from (.51, .72, .59) of the unsupervised method, over 2 million unlabeled words. Considering the knowledge-poor strategies we employed, and the fact that we did nothing to reveal the structure in compounds, this is quite striking, and shows us more avenues to move toward unsupervised segmentation. (1,838 words, out of 5,010 are either roots or compounds, which seems to be a representative percentage). We also note that we get .79 f-measure of correct segmentation into morphemes, i.e. we deliver the morphs, without morphological analysis, not just the overall tag for the word. What manual segmentation provides is syntactic and semantic disambiguation in an indirect way, hence some semantic-phonological cues (such as intonation, stress) and some limited syntactic knowledge (e.g. for compounds), are next targets we want to address. Our manual segmentation is going to be made publicly available at www.LcsL.metu.edu.tr/share. It will in time

grow up to a size of 45,000 words. We plan to take on MorphoChallenge data after this level of supervision.

5. References

Atalay, N. B., Oflazer, K., Say, B. (2003). The Annotation Process in the Turkish Treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora - LINC*, April 13-14, 2003, Budapest, Hungary.

Baroni, M., Matiasek, J., Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. *In Proceedings of the Workshop on Morphological and Phonological Learning at ACL '02*, pp. 11–20.

Bozsahin, C. (2002). The Combinatory Morphemic Lexicon. *Computational Linguistics*, *28*(2), pp. 145-186.

Cohen-Sygal, Y., Gerdemann, D., Wintner, S. (2003). Computational Implementation of Non-Concatenative Morphology. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop*, pp. 59-66, Budapest, Hungary.

Çöltekin, Ç. (2010). A Freely Available Morphological Analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta, May 2010.

Creutz, M., Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, pp. 106-113.

Goldwater, S. J. (2007). Nonparametric Bayesian models of lexical acquisition. Ph.D. thesis. Brown University, Providence, RI, USA.

Güngör, T. (2003). Lexical and Morphological Statistics for Turkish. In *Proceedings of TAINN* 2003, pp. 409-412.

Hajic, J., Hladka, B. (1997). Tagging of inflective languages: a comparison. In *Proceedings of ANLP'97*, Washington, DC, pp. 136–143. ACL.

Hakkani-Tür, D. Z., Oflazer, K., Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities, 36*(4), pp. 381–410.

Hankamer, J. (1986). Finite state morphology and left to right phonology. In *Proceedings of the West Coast Conference on Formal Linguistics*, Volume 5. Stanford Linguistic Association.

Kiraz, G. A. (2002). *Computational Nonlinear Morphology, with Emphasis on Semitic Languages.* Cambridge, U.K.: Cambridge University Press.

Kohonen, O., Virpioja, S., Lagus, K. (2010). Semi-supervised learning of concatenative morphology. In *Proceedings of the Eleventh Meeting of the ACL* Special Interest Group on Computational Phonology and Morphology (SIGMORPHON 2010), Uppsala, pp. 78–86.

Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form recognition and generation. Ph.D. Thesis. University of Helsinki.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2).

Oflazer, K., Say, S., Hakkani-Tür, D. Z., Tür, G. (2003) Building a Turkish Treebank, Invited chapter in A. Abeille (Ed.), *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers.

Sak, H., Güngör, T., Saraçlar, M. (2011). Resources for Turkish Morphological Processing. *Language Resources and Evaluation, 45*(2), pp. 249–261.

Say, B., Zeyrek, D., Oflazer, K., Özge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. In *Proceedings of the Eleventh International Conference of Turkish Linguistics*.

Yarowsky, D., Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the ACL*, pp. 207–216.

Yatbaz, M. A., Yuret, D. (2009). Unsupervised Morphological Disambiguation using Statistical Language Models. In *Proceedings of the NIPS 2009 Workshop on Grammar Induction, Representation of Language and Language Learning.*

Yatbaz, M. A., Yuret, D. (2010). Unsupervised Part of Speech Tagging Using Unambiguous Substitutes from a Statistical Language Model. In *Proceedings of the Coling 2010*, Beijing, pp. 1391–1398.

A Platform for Creating Multimodal and Multilingual Spoken Corpora for Turkic Languages: Insights from the Spoken Turkish Corpus

Şükriye Ruhi[a], Kerem Eryılmaz[a], M. Güneş. C. Acar[b]
[a] Middle East Technical University, [b] Ankara University
[a]Dumlupınar Blvd., No. 1, Çankaya, 06800 Ankara, Turkey; Ankara Üniversitesi İletişim Fakültesi, 06590 Cebeci Ankara, Turkey
E-mail: sukruh@metu.edu.tr, keryilmaz@gmail.com, acargunes@gmail.com

Abstract

Based on insights gained from the corpus design and corpus management work involved in the compilation of the Spoken Turkish Corpus (STC), this paper addresses the possibility of developing sustainable, comparable, multimodal spoken corpora for facilitating comparative studies on Turkic Languages, with the capacities of a digital platform that incorporates EXMARaLDA software suite and a web-based corpus management system (STC-CMS), which together provide an interoperable system that can be customized for the creation of spoken and written corpora. Section 2 highlights the significance of multimodal corpus resources for comparative research and the development of technologies, and describes the implementation in STC, especially focusing on its metadata parameters and the flexibility of its transcription tools for representing cross-linguistic variation. Section 3 addresses the issue of developing common infrastructure for corpus compilation that can facilitate data transfer between resources. The paper concludes with a brief discussion on the challenge for creating comparable spoken corpora for the Turkic languages in regard to orthographic systems.

1. Introduction

Computerized language resources support the development of tools for automatic processing, machine translation, recall and retrieval of information from texts, to cite but a few of their computational functions. A further significant aim of developing such language resources is their capacity to support traditional language studies (e.g. language education, cultural studies, and comparative linguistics, to name a few). Within the context of speech and spoken corpora, while Turkish appears to enjoy a better representation within such resources compared to other Turkic languages (e.g. the Turkish Speecon Database, Salor, Çiloğlu and Demirekler, 2005; METU Turkish Corpus, Say et al., 2002), there is a paucity of computerized resources that can address the needs of both linguistic research and language technologies regarding Turkic languages as a whole linguistic group. Commenting on best practices regarding the creation of digital language resources, Nenadić (2004), underscores the importance of developing standardized, flexible, and multimodal resources that are "open to multilingual integration", and that can be used to "integrate language resources" through "transfer of competence and know-how". The purpose of this paper is to make a modest contribution in regard to the corpus design and technical infrastructure for multimodal and multilingual resources for spoken corpora of the Turkic languages. Based on work and tools developed in the Spoken Turkish Corpus (STC), which is designed to be a one-million-word, web-based corpus of spoken Turkish discourse in its initial stage, the paper highlights the significance of open source, flexible and interoperable corpus compilation tools that are accessible to (non-)expert corpus compilation and annotation teams.

2. Designing Comparable Corpora

While comparability and standardization in spoken corpus design and metadata parameters –at least at the generic level– is essential for all kinds of (digital) language resources employed in cross-linguistic studies, it is all the more essential if multilingual resources are being designed with a view of aiding both the construction of further tools in the computational sciences and research in linguistics (see, the International Corpus of English, Greenbaum and Nelson (1996) for the latter purpose). In this paper, we highlight issues related to text classification, metadata features that can render corpus content more visible to end users, and text annotation, and describe how these have been approached in STC.

2.1 Text classification and metadata

Text classification is a crucial parameter in carrying out comparative research; however, classification in spoken corpora is still a standing debate (Lee, 2001) owing to the often dynamic and fluid purpose of interaction. A common procedure has been to categorize communications according to speaker features and speech genres by attending to what may be described as a mixture of discourse goal and discourse topic (e.g. the British National Corpus). More recent spoken corpora have employed double axes in classification. The Cambridge and Nottingham Corpus of Discourse in English (CANCODE), for example, distinguishes between goal of interaction and speaker relationships. Speaker relations are classified as: transactional professional, pedagogical, socializing and intimate. Goal types are distinguished as: provision of information, collaborative tasks and collaborative ideas (McCarthy 1998: 9-12). Family members cooking together, for example, would be intimate and collaborative task, while family members talking about family issues is intimate and collaborative ideas.

A problem with this type of classification is the rigid boundaries it draws for goal classification and in some cases for speaker relations. One can easily imagine, for

instance, that an intimate conversation may be both task and idea oriented at different times within the same communicative setting. If one were to compare CANCODE with BNC then, there would be difficulty in comparing the intimate and socializing categories with the leisure category in BNC. A further deficiency from the perspective of multilingual corpora is that languages and cultures may conceptualize social activities in different ways, so that the generic, double axes method might not reflect the true nature of the interaction at the more granular level and in a way that would correspond to the experience of the participants in the setting. In other words, the broad classificatory parameters in most present-day corpora do not make the socio-cultural situatedness of communication visible to researchers, so that searching for relevant tokens of sociolinguistic and pragmalinguistic phenomena needs to rely *per force* on in-depth qualitative investigation of the resource (Virtanen, 2009).

One way of handling the representation of the fluidity in interaction is to classify files first on the level of social relations and situational settings, and then incorporate second level metadata for the kind of social and discoursal activities that the participants engage in. Another dimension at this level is to indicate discourse topics (i.e. content) in the file metadata. This approach is essentially a method of combining generic metadata for text classification with more specific annotation at the pragmatics level, and is in line with research that emphasizes the notion of activity types in discourse and corpus annotation (e.g. Gu, 2009).

2.2 A pragmatically informed metadata

While STC has taken into consideration the text classification and other metadata parameters proposed in standardization schemes (e.g. the ISLE Meta Data Initiative – IMDI; Dublin Core – DC) and features in other spoken corpora (e.g. Spoken Dutch Corpus; see Oostdijk, 2000) (see, Ruhi et al. (2010b) for details on the classificatory and descriptive categories in STC), to attain text descriptors that are more fine-grained at the pragmatic level, STC implements a two-layered scheme regarding text type and discourse content.

On the first level, texts are classified according to speaker relations and the major social activity type. The domains for speaker relations are: family/relatives, friend, family-friend, educational, service encounter, workplace, media discourse, legal, political, public, research, brief encounter, and unclassified conversations (Ruhi et al., 2010b). These domains are then sub-classified according to activities. The class of workplace discourse, for instance, includes meetings, workplace cultural events (e.g. parties), business appointments, business interviews, business dinners, shoptalk, telephone conversations, and chats.

The second layer of metadata annotation is implemented at the corpus assignment stage and is checked in the cyclic steps in the transcription stage in STC (see, Ruhi et al. 2011). This layer involves the annotation of speech acts based on Searle (1976) (e.g. offers and requests), on the one hand, and on the other hand, the annotation of conversational topics (e.g. child care), speech events (e.g. troubles talk), and ongoing activities (e.g. cooking) –all encoded under the super metadata category, Topic, in the current state of STC. Speech act and Topic annotation are thus two further metadata parameters in STC. An overview of the corpus in terms of text categories, distribution of gender and age is available at the web site of STC and in its demo version. Figure 1 displays part of the metadata for a file in the service encounter domain, where the main activity and speech event is a sales transaction, i.e. buying a ticket.

Date recorded	2009-04-19T18:45:00
Domain	Service encounters
Duration	69
Genre	Shopping
Physical space	Travel agency
Relations	MEH000222 is service provider of MED000112.
Speech acts	Greetings, Requests, Compliance (as a response to a request), Offering, Leaves taking, Well wishes/congratulations, Thanking
Topics	Bilet alma
project-name	ODT-STD
transcription-convention	ODT-STD-HIAT
transcription-name	047_090419_00077

Figure 1. Partial metadata for a file in STC

Of note in the context of this paper is that these parameters are also being used in the construction of Spoken Turkish Cypriot Dialect Corpus (STCDC) with success. The addition of pragmatic and content metadata allows for retrieval of a variety of fairly formulaic language use and other conversational phenomena that are of interest both for cross- and intra-linguistic language variation (e.g. speech act variation and socio-phonetic stylistic shifts; Ruhi, forthcoming; Ruhi et al., 2011) and for the development of NLP applications (e.g. computational lexicons).

2.3 Annotation software for multilingual spoken corpora

Achieving standardization in representing spoken language is obviously crucial in multilingual corpora and involves a variety of corpus annotation and format design decisions that need to address both language specific features and the fact that the resource will be used for comparative purposes in different technological environments, each with their own traditions of processing data. In regard to creating multilingual spoken corpora for Turkic languages, a desirable annotation scheme would be to allow for the possibility of interlinear translation. A further desirable capacity for language representation would be construct multimodal corpora that can allow researchers to consult audio- and video files in a time-aligned manner with transcription files.

Today, spoken corpus builders have at their disposal quite a wide range of corpus annotation tools (e.g. ELAN, EXMARaLDA and TASX). While each of these software suites have their strengths and weaknesses, STC has opted to use EXMARaLDA (Extensible Markup Language for Discourse Analysis) for the following reasons:

(1) it guarantees long-term sustainability, since it is an open source system of data models, formats and tools for the production and analysis of spoken language corpora, with various export options. Amongst these are HTML, PDF, and RTF. The TEI-conformant option based on P5 markup (Schmidt 2011) and XML-based EXMARaLDA formats, which ensure accessibility, long-term archivability and interoperability, are particularly noteworthy (Schmidt, 2004; see, Figure 6 in Section 3.1, from Ruhi et al. 2010a);

(2) it can operate with a number of widely used transcription conventions (e.g. CHAT and GAT), which allows annotations to be customized according to research foci, and has built-in Praat analysis and IPA annotation files;

(3) it allows for both built-in speech annotation on a number of levels (e.g. for background events, translations, annotation of linguistic variation; see Figures 3 and 4) and stand-off annotation of the linguistic and socio-pragmatic features of communications (see, Ruhi, forthcoming; Ruhi et al., 2011); and perhaps as important as the reasons above,

(3) it has been localized for Turkish, and has been used for transcribing texts in Urum (Skopeteas and Moisidi, 2011).

STC has opted to employ EXMARaLDA tools with an adapted form of HIAT in its transcriptions, and has published a full documentation of the STC conventions, which allow for dialectal variation representation as well (see, Ruhi et al., 2010c). Figure 2 presents a snapshot from a file with video support, and illustrates a few of the annotations implemented in STC, including dialectal pronunciation.

Figure 2. A snapshot from a STC file

The same corpus annotation software and conventions are currently being employed in STCDC. The only difference between STC and STCDC is that the former primarily employs standard written orthography in representing utterances, while the latter incorporates conventionalized dialectal renderings in the utterance representation tier. Thus STC and STCDC are mirror images of each other in this respect (see, Figures 3 and 4 for the representation of k-g variation in the two corpora).

BUR000030 [v]	rdesin?	((0.6)) geç içeri.	_geç.	el.
			_g	
MUS000031 [v]			şunu al	kız!
MUS000031 [c]				gl
IND000002 [v]				
[nn]	((loud noise))			

Figure 3. k-g variation in STC

..*	5*
BUR000002 [v]	((0,8s)) ((OKA)) abiler **daşındıydı** burdan
BUR000002 [c]	*taşındıydı*
FIK000003 [v]	
FIK000003 [c]	
[nn]	*background))*

Figure 4. k-g variation in STCDC

Although, STC and STCDC do not provide in-built morphological analyses and translations, it is possible to add such annotation into the transcription files, as can be seen in the example from Urum below (Skopeteas and Moisidi, 2011):[1]

bizum xalx gyaldi kavkaza vasemnadsati vektya

our people came caucasus_to eighteenth century

'Our people came to the Caucasus in the eighteenth century.'

Figure 5. Morphological annotation and translation in the Urum narrative corpus

The flexibility that EXMARaLDA allows in regard to working with different stylesheets and export options, we would argue, make the system particularly compatible with multilingual spoken corpora building.[2] Indeed, quite a few of the ongoing or completed corpus projects that use EXMARaLDA are devoted to multilingual compilations (see, http://exmaralda.org/).

3. Corpus Management Across Diverse Populations and Long Distances: STC-CMS

Spoken corpus building, especially of the general corpus kind, relies on the support of a range of expert and non-experts: linguists with specializations conversation

[1] STC is currently piloting a web-based system for morphological analysis, using TRmorph (Çöltekin, 2010).

[2] EXMARaLDA can also be used for written corpus compilation.

analytic techniques, transcribers, IT infrastructure experts and programmers, institutional managers, and the general public in the curating recordings. This often translates into a 'working' population, who may be geographically distant, and who may not necessarily share each other's metalanguage. Thus developing a workflow and a corpus management system that responds to the diverse needs of such groups is a major challenge.

In STC, we opted for developing a web-based corpus management system, namely, STC-CMS, which prioritizes the issues raised above and takes interoperability to be a crucial feature for corpora. A detailed description of the workflow in STC is available in Acar and Eryılmaz (2010). In this section, we concentrate on the capabilities of STC-CMS.[3]

STC-CMS is a web-based system, developed in the project to make the management and the monitoring of corpus production easy, transparent and consistent. The system aims to attain maximum automation and validation, as well as a clearly defined, traceable workflow, which allows for monitoring the design parameters of the corpus and the progress of the workflows, and for maintaining consistency in producing the resource (see, Figure 6). As STC employs EXMARaLDA, a core function of STC-CMS is to achieve integration with its tools. STC-CMS performs this by generating EXMARaLDA compatible transcription and corpus metadata files.

The system enables smooth control of the media and metadata files through a web interface and a relational (MySQL) database for metadata. Contributors submit recordings and metadata through the web forms, where they are validated and added to the database. At this stage, STC-CMS generates the EXMARaLDA compatible transcription files, which makes it possible to use EXMARaLDA tools and formats in STC. When a transcription file is submitted, it is checked into an SVN system for backup measures.

Using various file and data formats, STC tries to minimize obsolescence. Amongst its notable features, the system allows any subset of the corpus to be defined and published, using EXMARaLDA libraries, through a password restricted web site, where anyone with a web browser may access the corpus. As noted in Section 2.3 too, the system harnesses EXMARaLDA's capabilities for exporting to different transcription systems like Praat, ELAN, and TASX Annotator (see Schmidt (2005) for a detailed description of the relation between EXMARaLDA and TEI formats; and Schmidt (2011) for TEI-conformant transcription options).

While STC-CMS was constructed for the project, it aims to function as an open source project for further enhancement of its capacities and use by resource producers who may wish to develop their own corpora.

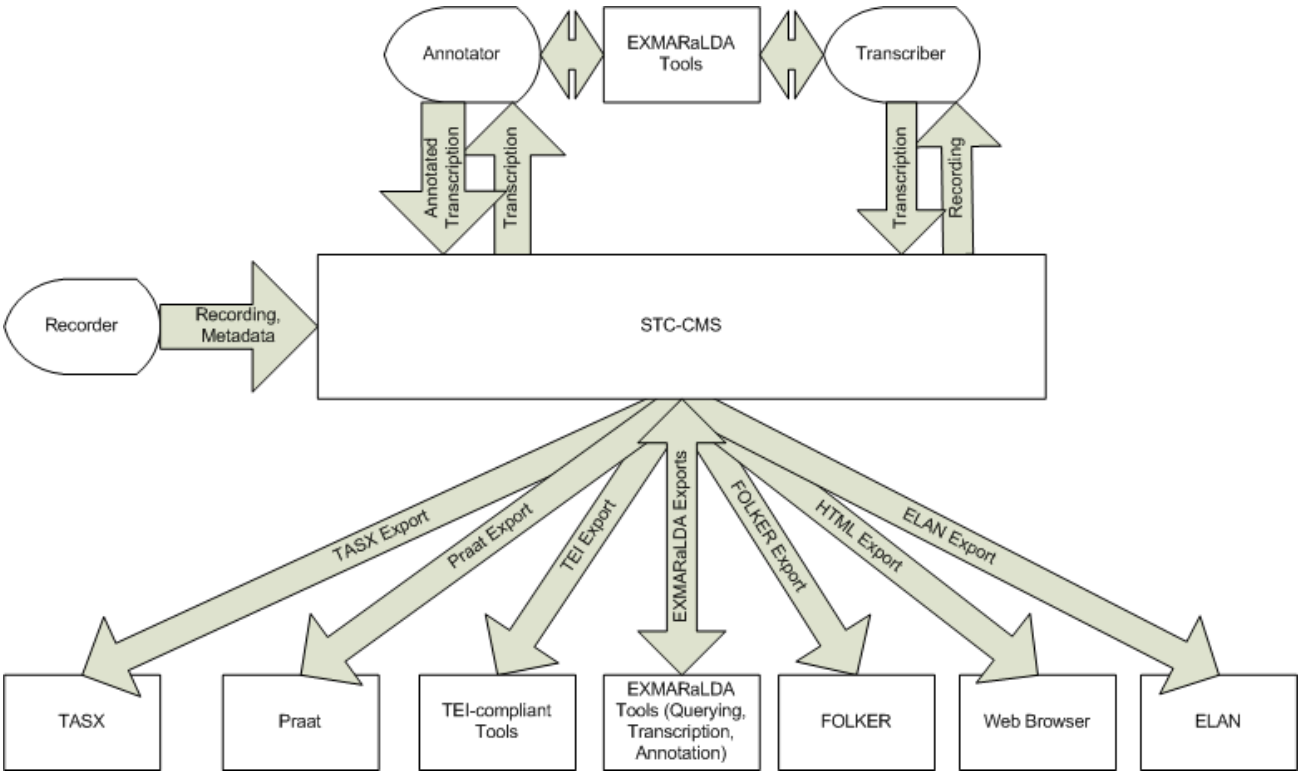

Figure 6. STC workflow and interoperability

[3] The description of STC-CMS in this section is largely based on Ruhi et al. (2010a).

Given that STC-CMS is web-based and that, together with EXMARaLDA software, it can be customized for individual corpus projects (especially in regard to the representation of linguistic, paralinguistic and non-verbal data), we suggest that it offers a viable system for multilingual corpora construction.

4. Issues and Challenges for Creating Multilingual Corpora

Creating a spoken corpus is a labor intensive and expensive enterprise. Thus one feature that one would expect of a spoken corpus is that it achieves maximum usability. While there are certainly several aspects to consider in developing a common infrastructure for creating multilingual corpora, in this paper we highlight the issue of parallelism in corpus metadata, speech representation and localization in digital systems for Turkic spoken corpora.

Multilingual corpora obviously need to achieve a minimum of common metadata features, especially concerning speaker attributes. Given that multilingualism and mobility characterize large bodies of populations around the globe today, we would suggest that language profiles, place(s) and length of residence, and possibly shifting professional profiles over time are amongst the features that are highly significant in multilingual corpora. To our knowledge, despite certain divergences regarding granularity in speaker and situational variables, metadata systems deployed in spoken corpora consider speakers within a snapshot in time in log files. That is to say, features that are valid at the time of recording are entered into metadata. Speakers in real life, however, have a life history such that their social roles in situated interaction as reflected in the metadata are only minimally representative of themselves as social agents. Language profiles and residence can be fine-tuned in metadata today (see, e.g., IMDI). However, a perspective that is largely ignored in current metadata practices is that speakers are 'social agents in a community', with multiple social roles that may show change over time yet impact the situated communication. This limitation is something that needs to be addressed by incorporating greater flexibility in standardization practices.

Regarding speech context features, there are likely to be speech genres and situations of communication that may not be captured if metadata practice does not allow for flexible category identification. Indeed, this is not an issue that bedevils multilingual corpora only. For example, mediated communication today is accomplished not solely through telephones today but also through various forms of video conferencing. Thus spoken communication speech genres need to be updated in these respects.

To address the issues noted above, STC-CMS has adopted a flexible metadata design that allows for the introduction of descriptive and classificatory categories for speaker and speech context features. In the context of multilingual corpora, it is possible to implement versions of STC-CMS metadata designs to cater for local needs. However, extended life histories in regard to social roles and multiple professional profiling is not an issue that we have been able to tackle yet.

Turning to the issues concerning orthography, systems vary amongst and within the same language in the Turkic languages group. On the one hand, there are languages such as Kazakh and Uyghur, which are arguably two instances that show the greatest regional variation. Kazakh employs three systems: Cyrillic, Latin and Arabic, while Uyghur employs these and also a pinyin system. On the other hand, there are cases such as Uzbek, which is moving from the Cyrillic to the Latin system in Uzbekistan.[4]

Phonetic representation could be considered an option where there is such diversity. This, however, would restrict end user accessibility, even if we set aside the problems it would raise regarding finding and training transcribers conversant with phonetic transcription. Furthermore, phonetic transcription increases the degree of interpretative annotation involved in corpus construction, which might not be the best solution given that corpus annotations are expected to be as consensual as possible (Leech, 1993).

In regard to the capacities of EXMARaLDA for transcription, the Cyrillic system is enabled through the virtual keyboard, while transliteration is the option for Arabic script. In cases where there are wide divergences or lack of an established orthographic system, transliteration might be an option to pursue, but we doubt that all end users would welcome such a solution. These are issues, naturally, where ultimate decisions depend on the purpose of creating the digital resource. Despite the presence of this issue, we end with a positive note on the technical side, and that is that localization within the tools described in this paper is possible.

5. Acknowledgements

STC was supported by TÜBİTAK 108K285 between 2008-2010 and is currently being partially supported through METU BAP-05-03-2011-001. We gratefully acknowledge the ongoing support of Thomas Schmidt and Kai Wörner.

6. References

Acar, G. C., Eryılmaz, K. (2010). Sözlü Derlem için Web Tabanlı Yönetim Sistemi. In *24. Ulusal Dilbilim Kurultayı Bildiri Kitabı*. Ankara: ODTÜ, Yabancı Diller Eğitimi Bölümü, pp.437--443.

BNC. www.natcorp.ox.ac.uk

Çöltekin, Ç. (2010). A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC2010), Valletta, Malta, May 2010.

[4] Data retrieved from http://aatturkic.org/

ELAN. http://www.mpi.nl/tools/elan/html

EXMARaLDA. http://exmaralda.org/

Greenbaum, S., Nelson, G. (1996). The International Corpus of English (ICE) Project. *World Englishes* 15(1), pp. 3--15.

Gu, Y. (2009). From real-life situated discourse to video-stream data-mining: An argument for agent-oriented modeling for multimodal corpus compilation. *International Journal of Corpus Linguistics*, 14(4), 433--466.

ISLE Meta Data Initiative. http://www.mpi.nl/IMDI/

Lee, D. YW. (2001). Genres, registers, text types, domains, and styles: clarifyıng the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3), 37--72.

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*, 8(4), pp. 275--281.

McCarthy, M. (1998). *Spoken languages and Applied Linguistics*. Cambridge: Cambridge University Press.

Nenadić, G. (2004). Creating digital language resources. Review of the National Centre for Digitisation, 5, pp. 19--30. Pre-print retrieved from http://personalpages.manchester.ac.uk/staff/G.Nenadic/papers%5CNCD_2004_Nenadic.pdf

Oostdijk, N. (2000). Meta-Data in the Spoken Dutch Corpus Project. LREC 2000 Workshop, Athens. http://www.mpi.nl/IMDI/documents/2000%20LREC/oostdijk_paper.pdf

Ruhi, Ş., forthcoming. Corpus linguistic approaches to (im)politeness: Corpus metadata features and annotation parameters in spoken corpora. In D. Z. Kádár, E. Németh, & K. Bibok (Eds.), *Politeness: Interfaces*. London: Equinox.

Ruhi, Ş., Eröz-Tuğa, B., Hatipoğlu, Ç., Işık-Güler, H., Acar, M. G. C., Eryılmaz, K., Can, H., Karakaş, Ö., and Çokal Karadaş, D. (2010a). Sustaining a corpus for spoken Turkish discourse: Accessibility and corpus management issues. In *Proceedings of the LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*. Paris: ELRA, 44-47. http://www.lrec-conf.org/proceedings/lrec2010/ workshops/W20. pdf# page=52

Ruhi, Ş., Işık-Güler, H., Hatipoğlu, Ç., Eröz-Tuğa, B., Çokal Karadaş, D. (2010b). Achieving representativeness through the parameters of spoken language and discursive features: the case of the Spoken Turkish Corpus. In I. Moskowich-Spiegel Fandiño, B. Crespo García, , M. Inés Lareo, P. Lojo Sandino (Eds.), *Language Windowing through Corpora. Visualización del Lenguaje a Través de Corpus*. *Part II*. A Coruña: Universidade da Coruña, 789--799.

Ruhi, Ş., Hatipoğlu, Ç., Işık-Güler, H., Eröz-Tuğa, B. (2010c). *A Guideline for transcribing conversations for the construction of Spoken Turkish corpora using EXMARaLDA and HIAT.* ODTÜ-STD: Setmer Basımevi.

Ruhi, Ş., Schmidt, T., Wörner, K. and Eryılmaz, K. (2011). Annotating for precision and recall in speech act variation: The case of directives in the Spoken Turkish Corpus. In *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*. *Working Papers in Multilingualism* Folge B, 96, 203--206.

Salor, Ö., Çiloğlu, T., Demirekler, M. (2005). Turkish Speecon Database. http://catalog.elra.info/product_info.php?products_id=800

Say, B., Zeyrek, D., Oflazer, K., and Özge, U. (2004). Development of a Corpus and a Treebank for Present-day Written Turkish. In *Current Research in Turkish Linguistics: Proceedings of the 11th International Conference on Turkish Linguistics*. Gazimagosa: Eastern Mediterranean University Press, pp. 183--192.

Schmidt, T. (2004). Transcribing and Annotating Spoken Language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris: ELRA. http://www1. uni-hamburg.de/exmaralda/Daten/4D-Literatur/Paper_LREC.pdf

Schmidt, T. (2005). Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. In *Arbeiten zur Mehrsprachigkeit, Folge B 62*.

Schmidt, T. (2010). Linguistic tool development between community practices and technology standards. In *Proceedings of the LREC Workshop Language Resource and Language Technology Standards – State of the Art, Emerging Needs, and Future Developments*. Valletta, Malta: European Language Resources Association (ELRA), 69--72.

Schmidt, T. (2011). A TEI-based approach to

standardising spoken language transcription. *Journal of the Text Encoding Initiative*, 1, pp. 1--22.

Skopeteas, S., Moisidi, V. (2011). *Urum Narrative Collection*. Urum Documentation Project. **http:// projects.turkmas.uoa.gr/urum/index.html**

Searle, J. (1976). A classification of illocutionary acts. *Language and Society* ,5, pp.1--23.

Spoken Turkish Corpus. http://std.metu.edu.tr/en/

Spoken Turkish Cypriot Dialect Corpus. http://corpus.ncc. metu.edu. tr/ ?lang=en

TASX. http://medien.informatik.fh-fulda.de/taskforce/ TASX-annotator

Virtanen, T. (2009). Corpora and discourse analysis. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook, Volume 2*. Berlin/New York: Walter de Gruyter, pp. 1043--1070.

.

Evaluation of Sentence Alignment Methods for English-Turkish Parallel Texts

Eray YILDIZ, A. Cüneyd TANTUĞ
İTÜ Computer and Informatics Faculty
Maslak, Istanbul, 34469
E-mail: yildizeray@hotmail.com.tr, tantug@itu.edu.tr

Abstract

In this work, we evaluate the performances of sentence alignment methods on aligning English-Turkish parallel texts. Three publicly available tools employing different strategies are tested in our study: a sentence length-based alignment method, a lexicon-based alignment method and a machine translation based alignment method. Experiments are carried out on a test dataset of parallel texts collected from web, mostly from newspapers. Due to the highly inflectional and derivational morphological structure of Turkish, we have incorporated stemming pre-processing step for the lexicon based tests. However, finding stems of Turkish wordforms requires a full morphological analysis and morphological disambiguation. So, as a simpler alternative stemming method, we suggest taking only *k*-characters of the wordforms as stems. Our experiments show that lexicon based methods with stemming performs best among all methods.

1. Introduction

For many natural process applications such as machine translation, cross-language information retrieval, word disambiguation having a parallel corpus is a very crucial and important initial step. A parallel corpus is comprised of aligned sentences that are translations of each other. Depending on the application type, words in these sentences can also be aligned for building a word level alignment. The process of constructing a parallel corpus has two main steps: collecting parallel texts which are translations of each other and the sentence alignment task in order to map sentences on the source side to their translations on the target side.

A number of automatic sentence alignment approaches have been proposed for sentence alignment problem. These approaches are based on different kind of knowledge levels like the length of sentences (number of characters or words), the amount of word correspondences (proper nouns, time stamps etc.), a bilingual lookup dictionary or an automatic machine translation (MT) system. In our study, we examine length-based, lexicon based and MT-based sentence alignment techniques on English-Turkish sentence alignment problem. The main motivation of this paper is having a comparative evaluation of publicly available sentence alignment tools with different approaches on Turkish-English sentence alignment problem. Building a parallel corpus or proposing a novel sentence alignment method is beyond the scope of this paper.

Section 2 gives brief information about related works while section 3 is devoted the explanations of the approaches and tools used in our evaluation. We mention the details of the experiments in Section 4 and present the experimental results in Section 5. The final section includes conclusions and discussions.

2. Related Work

Most of the sentence alignment approaches are based on sentence length comparisons and word correspondence checking, or combination of two. Initial works on sentence alignment are focused on sentence length only. Brown, Lai and Mercer (1991) have developed an algorithm based on sentence lengths by counting the number of words whereas Gale and Church's (1993) algorithm relies on the number of characters to calculate sentence length. The Gale and Church algorithm is still widely in use. As a latest example, it is used to align sentences in Europarl corpus (Senrich & Volk, 2010). The length based approaches work remarkably well on language pairs with high correlation in sentence lengths, such as French and English. On the other hand, the performance of length based aligners decrease significantly for the language pairs with low length correlation such as Chinese and English (Ma, 2006). A number of studies, such as (Li et al., 1994) and (Melamed, 1997) try to develop robust methods based on the sentence location information. These approaches are called geometric sentence alignment (GSA) approach that use sentence pair location information for aligning sentences. Wu (1994), try to overcome the weaknesses of length based approaches by utilizing lexical information from translation lexicons, and/or through the identification of cognates (Ma, 2006). Ma's (2006) lexicon-based sentence alignment approach increases the robustness of the alignment by assigning greater weights to less frequent translated words. The basic idea of MT based sentence alignment approaches is using machine translations of a text and MT evaluation scores to calculate a similarity score to find reliable alignments (Senrich&Volk, 2010).

Taşçı et al. (2006) develops a sentence alignment method for Turkish-English parallel sentences based on combination of sentence lengths and locations. A collection of parallel texts from e-books, news articles, academic works and translation companies' documents are compiled in this study. A new sentence alignment approach is presented and test on this Turkish-English parallel corpus. Accuracy rates up to 96% on the document pairs that have similar paragraph counts are achieved. Unfortunately this aligner is not available as a tool for public access.

In the literature, several studies focused on the evaluation of parallel text alignment techniques. The ARCADE project is an evaluation exercise dedicated to two main tasks: producing a reference bilingual corpus, aligned at sentence level and evaluating several sentence alignment systems (Langlais et al., 1998). Caseli and Nunes (2003) evaluate several sentence alignment systems on Portuguese-English parallel sentence pairs. Lambert et al. (2010) evaluate several length based, MT based and dictionary based sentence alignment systems on Urdu-English and French-English language pairs.

3. Methods

Three methods with different strategies are tested in our study: a sentence length-based alignment method, a lexicon-based alignment method and a MT based alignment method.

Sentence alignment methods based on sentence lengths rely on number of words or characters in the sentences on both sides. In our experiments, we have used Bilingual Sentence Aligner (BSA) as a length-based sentence aligner (Moore, 2002). BSA exploits not only sentence length for alignments but also word correspondences like proper nouns and date-time expressions. Moore's method is similar to Wu's (1994) method in that it uses both sentence length and lexical correspondences to derive the final alignment, but BSA doesn't require a lexicon. It is a simple and fast method like other systems based on sentence lengths. BSA has two pass algorithms. The first alignment subsequently serves as the training data for a translation model, which is then used in a complex similarity score calculation. Next, the algorithm works IBM-1 translation model to produce an alignment based both on sentence length and word correspondences (Moore, 2002). BSA only needs source and target text as input and does not necessitate any dictionaries. BSA can generate only 1-1 alignments.

Lexicon based sentence alignment methods makes use of an electronic bilingual dictionary for aligning sentences. We have used Champollion Tool Kit (CTK) as a lexicon based sentence aligner (Ma, 2006). CTK was initially developed for aligning Chinese-English parallel text. It was later ported to other language pairs, including Arabic–English and Hindi–English. CTK differs from other lexicon based sentence aligners in assigning weights to translated words. The weights are calculated with TF-IDF weighting method. While calculating the similarity scores, CTK penalizes the alignments other than 1-1 alignments. Also, sentences with a mismatching length are also penalized (Ma, 2006).

MT based sentence alignment methods does not align source and target side texts directly. In fact, these methods try to align target side texts with the translations of source side sentences obtained by a MT system. It is noteworthy that MT-based aligners work on sentences both in the same language. For example, in order to align English-Turkish sentences, the source (English) side sentences must be translated to Turkish with a MT tool. Then MT-based sentence aligner can align Turkish side sentences with the MT outputs, which are also in Turkish. For our study, we select the BleuAlign (Senrich&Volk, 2010) for a MT-based sentence aligner. Actually, BLEU (Papineni et al., 2002) has been developed as an automatic metric to measure the translation quality of MT systems by comparing the system output with one or more reference translations. This is done by measuring the token n-gram precision of the system translation for all n-gram levels up to 4, and combining the n-gram precisions using the geometric mean. In the first step, BleuAlign gives the target text and the MT outputs of the source text to BLEU and BLEU returns a score for similarity. The anchor points are identified in this step using this similarity score. In next step, the sentences between these anchor points are either aligned using BLEU-based heuristics or the length-based algorithm by Gale and Church (Gale, Church, 1993). In BleuAlign, the uni-grams of the words are used instead of 4 gram since BLEU scores are too small when 4 grams used. In our experiments, we used Google Translate [1] as the MT system required by MT-based methods.

4. Experiments

4.1 Test Dataset

The documents in the test data are collected from the web, commonly from bilingual news sites and bilingual 'about us' pages of universities and other institutions. The dataset contains 30 English–Turkish page pairs. The total number of sentences in the data set is 1035 and 1055 on English side and on Turkish side respectively [2]. The number of sentences in the test data set seems reasonable since most of the previous studies use test data sets having less than 1000 sentences (Taşçı et al., 2006) (Senrich&Volk, 2010).

4.2 Pre-processing

In order to build a golden standard data set for sentence alignment performance comparison, we aligned the sentences in the dataset manually. The alignments are denoted in the alignment files with the following output style:

```
    1 <=> 1      (1)
  2,3 <=> 2      (2)
```

(1) means that first sentence of source text is aligned with the first sentence of the target text. In the (2) alignment, the second and third sentences of the source text are aligned with the second sentence of target text. We obtained 947 alignments from 1035 English and 1055 Turkish sentences. 869 of them are 1-1 (one sentence to one sentence) and 78 are 1-N (one-to-many) and M-N (many to many).

[1] http://translate.google.com

[2] The test data set and the detailed information about the documents in the test data is avaible from http://ddi.ce.itu.edu.tr/resources/engtur_aligned.zip

English : It has advanced information systems and communication technology **facilities.**

Turkish : Gelişmiş bilgi ve iletişim teknolojileri **tesislerine** sahiptir.

Figure 1: Word matching problem for lexicon-based methods

4.3 Dictionary

Although CTK has Arabic-English, Chinese-English and Indian-English dictionaries, it does not involve an English-Turkish dictionary. For our experiments, we obtained and used an electronic English-Turkish dictionary which contains Turkish equivalents of 88.824 English words.

4.4 Stemming

CTK applies light stemmers to the sentences in English, Arabic and Chinese languages. The stemmer is used to normalize the words to their dictionary forms so that the number of lexical matches is maximized (Ma, 2006).

Lexicon based aligners utilizes bilingual dictionaries for finding the best match among possible sentences on the other side. For a successful dictionary lookup, the lemma form of a wordform must be searched in the dictionary. Similarly, lemma forms of the lookup results must be searched in the sentences on the other side. An example of word matching problem for lexicon-based methods is given in Figure 1. A lexicon based aligner performs a dictionary lookup for all words in the English sentence, and counts the number of matched equivalent words on the Turkish side. For example, the word "facilities" in the English sentence must be lemmatized and its lemma form "facility" must be searched in the dictionary. The bilingual dictionary entry for the word "facility" is shown below:

```
facility  (noun)    (1) tesisler
                    (2) kolaylıklar
                    (3) imkânlar
```

One can easily note that none of the equivalent Turkish words occurs in the target side Turkish sentence. Instead, the word "tesislerine", which is an inflected wordform of the dictionary lookup result "tesisler", occurs in target side sentence. This simple example shows the necessity of using a stemmer or lemmatizer for both dictionary lookup and target side lexical matching.

Whereas incorporating a stemmer in English is relatively easy, lemmatization in Turkish is not a trivial task. In order to find the lemma form of the wordform, a full morphological analysis and morphological disambiguation must be performed. We have used a Turkish morphological analyzer (Oflazer, 1994) and morphological disambiguation tool (Yuret&Ture, 2006) for finding the lemma forms of the wordforms in the Turkish sentences and dictionary entries.

This lemmatization process on Turkish side requires complex tools and poses overhead in the alignment process. Besides, the necessity of running a full morphological analysis for obtaining the lemma form is questionable since it is only used for matching. It may be more practical to use a light-weight stemmer. We suggest using a very naïve stemmer that assumes the first *k*-letters of a wordform as lemma form. Considering that the probability of having wordforms with the same first *k*-letters is not very high in a sentence, our naïve stemming method seems to be fairly good enough for our matching purposes.

In this study, we compare the performances of lexicon based sentence alignment on English-Turkish texts by employing both complex Turkish lemmatization process and our naïve stemmer which assumes the first *k*-letters as

Alignment Method		1-1			N-M			Overall		
		P	R	F1	P	R	F1	P	R	F1
Length Based Alignment		0.807	0.817	0,811	N/A	N/A	N/A	0.807	0.817	0,811
MT Based Alignment		0.940	0.903	0,921	0.494	0.564	0,527	0.897	0.875	0,885
Lexicon Based Alignment (without stemming)		0.905	0.787	0,843	0.244	0.448	0,330	0.800	0.759	0,779
Lexicon Based Alignment (with full stemming)		0.961	0.924	0,942	0.490	0.679	0,576	0.907	0.903	0,905
Lexicon Based Alignment (with naïve stemming)	k=2	0.958	0.884	0,920	0.362	0.641	0,481	0.871	0.864	0,867
	k=3	0.967	0.902	0,933	0.462	0.717	0,575	0.902	0.903	0,902
	k=4	0.970	0.936	0,952	0.509	0.717	0,604	0.916	0.918	0,917
	k=5	**0.970**	**0.940**	**0,954**	0.556	**0.756**	**0,648**	**0.924**	**0.925**	**0,924**
	k=6	0.970	0,940	0,954	**0,557**	0,753	0,647	0,924	0,924	0,924

Table 1: Sentence alignment performances

lemma. Additionally, in order to find the optimal value for *k*, we run multiple tests with *k = 2, 3, 4, 5* and *6*.

5. Experimental Results

The test results of different alignment approaches that are focused in this study are presented in Table 1. In this table, 1-1, N-M and overall sentence alignment performances are listed in separate columns. Since most of the NLP applications can only use 1-1 aligned sentences, having successful 1-1 alignments is usually more crucial in sentence alignment tasks. The performances of the different aligners are given in P (precision), R (Recall) and F-score metrics. The higher precision value is the more correct alignments whereas the higher recall value is the wider coverage of the actual alignments. Precision and recall values of BSA for N-M alignment are given as N/A because this tool is capable of producing 1-1 alignments only. Among the three alignment strategies, lexicon-based alignment with stemming performs best. The results for commonly used length-based alignment exhibits serious performance deterioration when compared to other two methods.

From the point of view of stemming effect in lexicon-based aligner, evaluation results reveal that substantial level (17%) of recall and considerable level (6%) of precision improvements are acquired by the help of stemming. In other terms, stemming process let the lexicon-based aligner produce a larger set of more accurate aligned sentences. Despite of the simplicity of our naïve stemming method, experimental results show that the performance of naïve stemming is almost same with the full stemming employing complex morphological analysis and disambiguation. The best performing results with naïve stemming are achieved with *k=5*. Moreover, for *k>3*, the lexicon-based method using naïve stemming performs even slightly better than the version using full-fledged stemming method. The main reason is that naïve stemming method allows matching of multiple word dictionary entries and typos.

6. Conclusions

This study evaluates three different alignment approaches for Turkish-English sentence alignment problem on a test data, which is comprised of 947 hand-aligned English-Turkish parallel sentences. As far as we know, this is the first effort on evaluating sentence alignment algorithms comparatively for English-Turkish case. Test results show that the lexicon-based sentence alignment method with stemming gets the best performance. As a novel contribution, we propose incorporating a light-weight naïve stemmer instead of the heavy stemming, i.e. stemming with full morphological analysis and morphological disambiguation. The lexicon-based aligner with our first *k*-letter based naïve stemming method succeeds to get the best alignment performance for *k=5*. As one of the three approaches focused in this study, MT-based aligner is also able to make alignments with high precision and recall values. Providing that a good MT system is accessible, MT-Based sentence aligners are able to produce reasonably good alignments.

Although its performance is low when compared to other two strategies, the length-based sentence alignment approach can also be preferred for the cases where acquiring an electronic dictionary or accessing a MT system is not feasible or too costly. We plan to use the results of this work in our efforts to build an English-Turkish parallel corpus from web.

7. References

Brown, P. F., Lai, J. C., Mercer , R. L. (1991). Aligning sentences in parallel corpora. In Proceedings of the 29th annual meeting on Association for Computational Linguistics, pp. 169–176.

Caseli, H.M., Nunes, M.G.V. (2003), "Evaluation of Sentence Alignment Methods on Portuguese-English Parallel Texts", Scientia. Vol. 14(2), pp. 223-238.

Gale, W. A., Church, K. W. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics., 19, 2, pp. 75–102.

Langlais P., Simard M., Veronis J. (1998). Methods and practical issues in evaluating alignment techniques. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL), Montr´eal, Canada,August.

Lambert P., Abdul-Rauf S., Fishel M., Noubours S., Sennrich R. (2010). Evaluation of sentence alignment systems. Fifth MT Marathon. Le Mans, France.

Li, W., Liu, T., Wang, Z., Li, S. (1994) Aligning bilingual corpora using sentences location information. Proc. of 3rd ACL SIGHAN Workshop, 141-147

Ma, X. (2006). Champollion: A Robust Parallel Text Sentence Aligner . LREC 2006: The Fifth International Conference on Language Resources and Evaluation.

Melamed, I.D. (1996) A geometric approach to mapping bitext correspondence. IRCS Technical Report 96-22, University of Pennsylvania

Moore, R.C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora . In AMTA ’02: *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pp. 135–144.

Oflazer, K. (1994). Two-level description of Turkish Morphology. Literary and Linguistic Computing.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method For Automatic Evaluation Of Machine Translation. In Proceedings of ACL.

Sennrich, R., Volk , M. (2010). MT-based Sentence Alignment for OCR-generated Parallel Texts . Proceedings of the 29th annual meeting on Association for Computational Linguistics, pp. 169—176.

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. ACL '94.

Yuret, D. Ture, F. (2006). Learning morphological disambiguation rules for Turkish Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

Taşçı, Ş., Güngör, A. M., Güngör, T. (2006). Compiling a Turkish-English Bilingual Corpus and Developing an Algorithm for Sentence Alignment. International Scientific Conference Computer Science’2006