Theory and Applications of Natural Language Processing
Edited volumes

Kemal Oflazer · Murat Saraçlar *Editors*

Turkish Natural Language Processing

Springer

Theory and Applications of Natural Language Processing

Series editors
Julia Hirschberg
Eduard Hovy
Mark Johnson

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

"Theory and Applications of Natural Language Processing" is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

* Downloadable on your PC, e-reader or iPad
* Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
* Available online within an extensive network of academic and corporate R&D libraries worldwide
* Never out of print thanks to innovative print-on-demand services
* Competitively priced print editions for eBook customers thanks to MyCopy service http://www.springer.com/librarians/e-content/mycopy

More information about this series at http://www.springer.com/series/8899

Kemal Oflazer • Murat Saraçlar
Editors

Turkish Natural Language Processing

Springer

Editors
Kemal Oflazer
Carnegie Mellon University Qatar
Doha-Education City, Qatar

Murat Saraçlar
Electrical and Electronic Engineering
Boğaziçi University
Istanbul-Bebek, Turkey

ISSN 2192-032X ISSN 2192-0338 (electronic)
Theory and Applications of Natural Language Processing
ISBN 978-3-319-90163-3 ISBN 978-3-319-90165-7 (eBook)
https://doi.org/10.1007/978-3-319-90165-7

Library of Congress Control Number: 2018948167

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Turkish has proved to be a very interesting language for natural language processing techniques and applications. There has been a significant amount of work on Turkish since the early 1990s on introducing and/or adapting fundamental techniques, compiling resources, and developing applications.

The idea for this book came after one of us gave an invited talk at the LREC Conference held in Istanbul, Turkey, in 2012. Since then, the authors and we have worked hard to bring this effort to fruition. This book brings together most of the work done on Turkish in the last 25 years or so. After a bird's-eye overview of relevant aspects of Turkish, it covers work on morphological processing and disambiguation, statistical language modeling, speech processing, named-entity recognition, dependency, and deep parsing. It then continues with statistical machine translation from English to Turkish and from Turkic languages to Turkish and sentiment analysis for Turkish, a topic that has recently been quite popular with the advent of social media. Finally, the book covers the most important natural language processing resources that have been developed for Turkish including the Turkish WordNet, the Turkish Treebank, Turkish National Corpus, and Turkish Discourse Bank.

We hope that this book helps other researchers in advancing the state of the art for Turkish and possibly other Turkic languages that share nontrivial similarities with Turkish.

Doha, Qatar — Kemal Oflazer
Istanbul, Turkey — Murat Saraçlar
July, 2017

Acknowledgements

The work presented in the following chapters has been supported by grants from various institutions: Early work was funded by a grant by NATO Science for Stability Program. The Turkish Scientific and Technological Research Council (TÜİTAK) has funded much of the more recent work in the last 15–20 years through many grants. Turkish Academy of Sciences (TUBA) provided support through its Outstanding Young Scientist Award program (TUBA-GEBIP). Boğaziçi University Research Fund (BU-BAP) also funded the language modeling and speech recognition research through grants. Additional funding was provided by the EU Framework 5 program.

A number of people have been very supportive of our work throughout these years: Lauri Karttunen has encouraged us and provided us with the Xerox Finite State Toolkit since the beginning, in order for us to build the very basic resources we needed. Other colleagues (then) at XRCE/PARC, notably Ken Beesley and Ron Kaplan, have supported our work by helping with the intricacies of the toolkit and by including our team in the ParGram Project. We thank them. The FSM Library and AMTools software packages developed by AT&T Labs-Research also have been essential in the early work on Turkish large vocabulary speech recognition. These were superseded by the OpenFST and Kaldi toolkits. We thank the authors of these toolkits and their organizations.

During these years, many graduate students and/or research assistants contributed to the work described in these chapters with their theses. Without their contributions, most of the work reported in the following chapters would not have been possible. We won't attempt to list them, lest we forget some. We thank them also.

Finally, we owe a lot to our families who supported us through these years. We cannot thank them enough.

Contents

About the Authors

Eşref Adalı received his B.Sc. and Ph.D. from Istanbul Technical University, Turkey, in 1971 and 1976, respectively. Currently, he is the Dean of the Faculty of Computer Engineering and Informatics, which he founded in 2010 after founding the Department of Computer Engineering in 1998. He was a Visiting Research Fellow at Case Western University, Cleveland, OH, USA, in 1977–1978; a Visiting Professor at Carleton University, Ottawa, Canada, in 1979; and a Visiting Professor at Akron University, Akron, OH, USA, in 1985. Adalı worked at the TÜBİTAK Marmara Research Center as the Founding Director of Informatics Group in 1990–1991. Although his more recent work has been on Turkish natural language processing, he has worked on microprocessors, system analysis, and design in the past and published two books on microcomputers and real-time systems.

Mustafa Aksan is Professor of Linguistics at Mersin University, Turkey, where he is also the Head of the Turkish Center for Corpus Studies. His main research interests are in corpus linguistics, lexical semantics, pragmatics, and morphology. He is the Principal Researcher in the construction of the *Turkish National Corpus* and also a coauthor of *A Frequency Dictionary of Turkish* (Routledge 2016). Recently, Aksan started a project on affix ordering in Turkish.

Yeşim Aksan is Professor of Linguistics at Mersin University, Turkey. Her main research interests are in corpus linguistics, lexical semantics, cognitive semantics, and pragmatics. She is the Project Director of the *Turkish National Corpus* and a coauthor of *A Frequency Dictionary of Turkish* (Routledge 2016). Previously, Aksan conducted corpus-based and corpus-driven studies on various aspects of Turkish.

Ebru Arısoy received her B.Sc., M.Sc., and Ph.D. degrees from the Electrical and Electronics Engineering Department, Boğaziçi University, Istanbul, Turkey, in 2002, 2004, and 2009, respectively. From 2010 to 2013, she worked as a postdoctoral researcher at IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. She then moved to IBM Turkey, where she had a key role in developing language modeling approaches for voice search and mobile dictation applications. Since 2014, she has worked as an Assistant Professor at the Electrical and Electronics

Engineering Department of the MEF University, Istanbul, Turkey. Her research interests include automatic speech recognition, statistical language modeling, as well as speech and language processing for educational technologies.

Orhan Bilgin is a translator, editor, and a dictionary publisher based in Istanbul, Turkey. He holds a B.A. in Economics and an M.A. in Cognitive Science from the Boğaziçi University, Istanbul, Turkey. His master's thesis was on frequency effects in the processing of morphologically complex nouns in Turkish. Between 2001 and 2004, he worked as a lexicographer on the Turkish team at the Sabancı University, Istanbul, Turkey, as part of the Balkanet project: an EU-funded project that designed and developed medium-sized word nets for six Balkan languages, including Turkish. Bilgin is the founder of zargan.com, an online English-Turkish dictionary that has been active since 2001, as well as a partner of Banguoglu Ltd., a translation company specializing in the translation of legal documents.

Cem Bozşahin works on the learning and projection of structure, including argument structure and constituent structure, at the intersection of computer science, linguistics, and philosophy. He holds a Ph.D. in Computer Science from the Arizona State University, Tempe, AZ, USA. He worked at the Ohio University (1990–1992) before permanently joining the Middle East Technical University, Ankara, Turkey, in 1992. He held visiting research assignments at the University of Edinburgh, Scotland, UK (2002–2017), at the Boğaziçi University, Istanbul, Turkey (2011), and at the University of Lisbon, Portugal (2015–2016).

Ruket Çakıcı is a lecturer at the Department of Computer Engineering of the Middle East Technical University, Ankara, Turkey. She received her Ph.D. from the School of Informatics of the University of Edinburgh, Scotland, UK, after receiving her M.Sc. and B.Sc. degrees in Computer Engineering from the Middle East Technical University. Her research interests are in empirical and data-driven methods for natural language processing, in particular (multilingual) morphological, syntactic, and semantic parsing; combinatory categorical grammar; automatic description generation for images and videos; and natural language generation.

Özlem Çetinoğlu is a postdoctoral researcher at the IMS, University of Stuttgart, Germany. She received her Ph.D. in Computer Science from the Sabancı University, Istanbul, Turkey, in 2009, where she developed a large-scale Turkish LFG grammar as part of her thesis in the context of the ParGram project. From 2009 to 2011, she worked at the CNGL, Dublin City University, Ireland, on automatically labeling English with deep syntactic information and on parsing noncanonical data. Çetinoğlu's research interests include deep grammars, statistical dependency parsing, morphologically rich languages, and code switching.

Işın Demirşahin holds a Ph.D. in Cognitive Science from the Middle East Technical University, Ankara, Turkey. Her research focuses on discourse structures in general and Turkish discourse and information structure in particular. Demirşahin currently works for Google UK Ltd. as a computational linguist and focuses on internationalizing end-to-end dialogue systems.

İlknur Durgar-El Kahlout is currently Chief Researcher at the TÜBİTAK-BİLGEM Research Center in Gebze, Turkey. She received her B.Sc. degree in Computer Engineering from the Başkent University, Ankara, Turkey, in 1997 and her M.Sc. and Ph.D. degrees in Computer Science and Engineering from the Sabancı University, Istanbul, Turkey, in 2003 and 2009 respectively. She worked as a postdoctoral fellow in the Quero project at LIMSI-CNRS, Orsay, France. Durgar-El Kahlout's current research interests are in natural language processing and statistical machine translation, especially of morphologically complex languages.

Gülşen Eryiğit obtained her M.Sc. and Ph.D. degrees from the Computer Engineering Department of the Istanbul Technical University, Turkey, in 2002 and 2007, respectively, where she is currently a faculty member. In addition, she is a founding member and coordinator of the Natural Language Processing Group and a member of the Learning from Big Data Group. In 2006, she worked as a Visiting Student Researcher at Växjö University, Sweden, in the first-ranked team in CoNLL shared tasks on multilingual dependency parsing. In 2007, Eryiğit won the Siemens Excellence Award for her Ph.D. thesis. Her current research focuses on natural language processing of Turkish on which she has authored and coauthored publications in prestigious journals and conferences. She represented Turkey in the CLARIN Project (EU 7th Framework Program, Common Language Resources and Technology Infrastructure) and in PARSEME (EU Cost Action, Parsing and Multiword Expressions). Recently, Eryiğit acted as the Principal Investigator of two research projects "Parsing Web 2.0 Sentences" (funded by EU Cost Action and TÜBİTAK) and "Turkish Mobile Personal Assistant" (funded by the Turkish Ministry of Science, Industry and Technology). Moreover, she is also the NLP coordinator of the interdisciplinary research project "A Signing Avatar System for Turkish-to-Turkish Sign Language Machine Translation" (funded by TÜBİTAK). Eryiğit also used her research to consult several domestic and international IT companies on Turkish natural language processing.

Gizem Gezici received her B.Sc. and M.Sc. degrees in Computer Science and Engineering from the Sabancı University, Istanbul, Turkey, in 2011 and 2013. She is currently a fourth-year Ph.D. student at the Sabancı University and is involved in a project in the emerging research area of *bias in search*. Gezici's research interests include sentiment analysis, machine learning, and data mining.

Dilek Zeynep Hakkani-Tür is a Research Scientist at Google. Prior to joining Google, she was a researcher at Microsoft Research (2010–2016), the International Computer Science Institute (2006–2010), and AT&T Labs-Research (2001–2005). She received her B.Sc. degree in Computer Engineering from the Middle East Technical University, Ankara, Turkey, in 1994, and her M.Sc. and Ph.D. degrees in Computer Engineering from the Bilkent University, Ankara, Turkey, in 1996 and 2000, respectively. Her research interests include natural language and speech processing, spoken dialogue systems, and machine learning for language processing. She coauthored more than 200 papers in natural language and speech processing and is the recipient of three Best Paper Awards for her work on active learning for dialogue systems from IEEE Signal Processing Society, ISCA, and EURASIP. She was an associate editor of *IEEE Transactions on Audio, Speech, and Language Processing* (2005–2008), a member of the IEEE Speech and Language Technical Committee (2009–2014), and an area editor for speech and language processing for Elsevier's *Digital Signal Processing Journal* and *IEEE Signal Processing Letters* (2011–2013). Since 2014, she has been a Fellow of IEEE and ISCA and currently serves on the ISCA Advisory Council (2015–2019). In addition, Hakkani-Tür was granted over 50 patents for her work.

Joakim Nivre is Professor of Computational Linguistics at Uppsala University, Sweden. He holds a Ph.D. in General Linguistics from the University of Gothenburg, Sweden, and a Ph.D. in Computer Science from Växjö University, Sweden. His research focuses on data-driven methods for natural language processing, in particular for syntactic and semantic analysis. Nivre is one of the main developers of the transition-based approach to syntactic dependency parsing, described in his book *Inductive Dependency Parsing* (Springer 2006) and implemented in the widely used MaltParser system. In addition to his current position of President of the Association for Computational Linguistics, he is one of the founders of the "Universal Dependencies" project, which aims to develop cross-linguistically consistent treebank annotation for many languages and currently involves over 50 languages and over 200 researchers around the world. As of July 2017, Nivre was cited more than 11,000 times and produced over 200 scientific publications.

Kemal Oflazer received his Ph.D. in Computer Science from Carnegie Mellon University in Pittsburgh, PA, USA, and his M.Sc. in Computer Science and B.Sc. in Electrical and Electronics Engineering from the Middle East Technical University, Ankara, Turkey. He is currently a faculty member at Carnegie Mellon University in Doha, Qatar, where he is also the Associate Dean for Research. He held visiting positions at the Computing Research Laboratory of the New Mexico State University, Las Cruces, USA, and at the Language Technologies Institute, Carnegie Mellon University. Prior to joining CMU-Qatar, he worked in the faculties of the Sabancı University in Istanbul, Turkey (2000–2008), and the Bilkent University in Ankara, Turkey (1989–2000). He has worked extensively on developing natural language processing techniques and resources for Turkish. Oflazer's current research interests are in statistical machine translation into morphologically complex languages, the use of NLP for language learning, and machine learning for computational morphology. In addition, he was a member of the editorial boards of *Computational Linguistics, Journal of Artificial Intelligence Research, Machine Translation*, and *Research on Language and Computation* and was a book review editor for *Natural Language Engineering*. Apart from having been a member of the Nomination and Advisory Boards for EACL, he served as the Program Co-chair for ACL 2005, an area chair for COLING 2000, EACL 2003, ACL 2004, ACL 2012, EMNLP 2013, and the Organization Committee Co-chair for EMNLP 2014. Currently, he is an editorial board member of both *Language Resources and Evaluation* and *Natural Language Engineering* journals and is a member of the advisory board for "SpringerBriefs in Natural Language Processing."

Murat Saraçlar received his B.Sc. degree in 1994 from the Electrical and Electronics Engineering Department of the Bilkent University, Ankara, Turkey; his M.S.E. degree in 1997; and Ph.D. degree in 2001 from the Electrical and Computer Engineering Department of the Johns Hopkins University, Baltimore, MD, USA. From 2000 to 2005, he was with the multimedia services department of the AT&T Labs Research. In 2005, he joined the Electrical and Electronic Engineering Department of the Boğaziçi University, Istanbul, Turkey, where he is currently Full Professor. He was a Visiting Research Scientist at Google Inc., New York, NY, USA (2011–2012) and an Academic Visitor at IBM T. J. Watson Research Center (2012–2013). Saraçlar was awarded the AT&T Labs Research Excellence Award in 2002, the Turkish Academy of Sciences Young Scientist (TUBA-GEBIP) Award in 2009, and the IBM Faculty Award in 2010. He published more than 100 articles in journals and conference proceedings. Furthermore, he served as an associate editor for *IEEE Signal Processing Letters* (2009–2012) and *IEEE Transactions on Audio, Speech, and Language Processing* (2012–2016). Having been editorial board member of *Language Resources and Evaluation* from 2012 to 2016, Saraçlar is currently an editorial board member of *Computer Speech and Language* as well as a member of the IEEE Signal Processing Society Speech and Language Technical Committee (2007–2009, 2015–2018).

Mark Steedman is Professor of Cognitive Science at the School of Informatics of the University of Edinburgh, Scotland, UK. Previously, he taught at the Department of Computer and Information Science of the University of Pennsylvania, Philadelphia, PA, USA, which he joined as Associate Professor in 1988. His Ph.D. on artificial intelligence is from the University of Edinburgh. He is a Fellow of the American Association for Artificial Intelligence, the British Academy, the Royal Society of Edinburgh, the Association for Computational Linguistics, the Cognitive Science Society and a member of the European Academy. Steedman's research interests cover issues in computational linguistics, artificial intelligence, computer science and cognitive science, including syntax and semantics of natural language, wide-coverage parsing and open-domain question answering, comprehension of natural language discourse by humans and by machine, grammar-based language modeling, natural language generation, and the semantics of intonation in spoken discourse. Much of his current NLP research addresses probabilistic parsing and robust semantics for question answering using the CCG grammar formalism, including the acquisition of language from paired sentences and meanings by child and machine. Some of his research also concerns the analysis of music by humans and machines. Steedman occasionally works with colleagues in computer animation where these theories are used to guide the graphical animation of speaking virtual or simulated autonomous human agents.

Umut Sulubacak is a research and teaching assistant at the Department of Computer Engineering of the Istanbul Technical University, Turkey. As part of his B.Sc. and M.Sc. studies, his research focused on the morphological and syntactic analysis of Turkish, using both rule-based and data-driven methods and optimizing morpho-syntactic annotation processes for Turkish dependency treebanks. He was involved in the construction of the Turkish treebank as part of the "Universal Dependencies" project and has remained an active contributor to the project ever since. In addition to his teaching responsibilities, he currently pursues his Ph.D. degree at the same institution with research in treebank linguistics and machine learning for Turkish language processing.

A. Cüneyd Tantuğ is currently Associate Professor at the Faculty of Computer and Informatics Engineering of the Istanbul Technical University, Turkey, where he completed his Ph.D. and has been a faculty member since 2009. His research areas include natural language processing, machine translation, and machine learning.

Gökhan Tür is a computer scientist focusing on human/machine conversational language understanding systems. He was awarded his Ph.D. in Computer Science from the Bilkent University, Ankara, Turkey, in 2000. Between 1997 and 1999, he was a Visiting Scholar at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA; then at the Johns Hopkins University, Baltimore, MD, USA; and at the Speech Lab of SRI, Menlo Park, CA, USA. He was at AT&T Research (2001–2006), working on pioneering conversational systems like "How May I Help You?" Later at SRI, he worked for the DARPA GALE and CALO projects (2006–2010). He was a founding member of the Microsoft Cortana team, focusing on deep learning methods (2010–2016), and was the Conversational Understanding Architect at the Apple Siri team (2014–2015). Tür is currently with the Deep Dialogue team at Google Research. Apart from frequent presentations at conferences, he coauthored more than 150 papers in journals and books. He is also a coeditor of the book *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech* (Wiley 2011) and of a special issue on Spoken Language Understanding of the journal *Speech Communication*. He is also the recipient of the Best Paper Award of *Speech Communication* by ISCA for 2004–2006 and by EURASIP for 2005–2006. Tür was the organizer of the HLT-NAACL 2007 Workshop on Spoken Dialog Technologies, the HLT-NAACL 2004, and the AAAI 2005 Workshops on Spoken Language Understanding. He also served as the area chair for spoken language processing for IEEE ICASSP conferences from 2007 to 2009 and IEEE ASRU Workshop in 2005, as the spoken dialog area chair for the HLT-NAACL conference in 2007, and as an organizer of the SLT Workshop in 2010. Having been a member of the IEEE Speech and Language Technical Committee (SLTC) (2006–2008) and the IEEE SPS Industrial Relations Committee (2013–2014) as well as an associate editor for the *IEEE Transactions on Audio, Speech, and Language Processing* (2010–2014) and *Multimedia Processing* (2014–2016) journals, Tür is currently a senior member of IEEE, ACL, and ISCA.

Berrin Yanıkoğlu is Professor of Computer Science at the Sabancı University, Istanbul, Turkey. She received a double major in Computer Science and Mathematics from the Boğaziçi University, Istanbul, Turkey, in 1988 and her Ph.D. in Computer Science from the Dartmouth College, Hanover, NH, USA, in 1993. Her research interests lie in the areas of machine learning, with applications to image and language understanding, currently focusing on multimodal deception detection, sentiment analysis, online handwriting recognition, and object recognition from photographs. Yanıkoğlu received an IBM Research Division award in 1996 and first place positions in several international signature verification competitions in collaboration with her students and colleagues.

Reyyan Yeniterzi received her B.Sc. and M.Sc. degrees in Computer Science and Engineering from the Sabancı University, Istanbul, Turkey, in 2007 and 2009, and her M.Sc. and Ph.D. degrees from the Language Technologies Institute of Carnegie Mellon University, Pittsburgh, PA, USA, in 2012 and 2015. Since 2015, she has worked as an Assistant Professor at the Computer Science Department of the Özyeğin University, Istanbul, Turkey. Previous to her various visiting research positions at the International Computer Science Institute (ICSI), Berkeley, CA, USA; at Vanderbilt University, TN, USA; and at Carnegie Mellon University in Doha, Qatar, Yeniterzi gained practical experience as an intern at Google. Her main research interests include natural language processing, text mining, social media analysis, information retrieval, search engines, and statistical machine translation.

Deniz Yuret is Associate Professor of Computer Engineering at the Koç University in Istanbul, Turkey, where he has worked at the Artificial Intelligence Laboratory since 2002. Previously, he was at the MIT AI Lab, Cambridge, MA, USA (1988–1999), and later cofounded Inquira, Inc., a company commercializing question answering technology (2000-2002). Yuret worked on supervised and unsupervised approaches to syntax, morphology, lexical semantics, and language acquisition. His most recent work is on grounded language acquisition and understanding.

Deniz Zeyrek is Professor of Cognitive Science and Director of the Informatics Institute at the Middle East University in Ankara, Turkey. She holds a Ph.D. in Linguistics from the Hacettepe University, Ankara, Turkey. Her broad research interests are in Turkish discourse and pragmatics, development of annotation schemes for recording discourse, and pragmatic phenomena on corpora. She contributed to the development of the METU Turkish Corpus, a corpus of modern written Turkish, and is the principal developer of Turkish Discourse Bank, a discourse corpus annotated in the PDTB style. Zeyrek's research mainly focuses on discourse relations and means of expressing discourse relations in Turkish and expands to discourse structure as revealed by discourse relations.

Chapter 1
Turkish and Its Challenges for Language and Speech Processing

Kemal Oflazer and Murat Saraçlar

Abstract We present a short survey and exposition of some of the important aspects of Turkish that have proved to be interesting and challenging for natural language and speech processing. Most of the challenges stem from the complex morphology of Turkish and how morphology interacts with syntax. Finally we provide a short overview of the major tools and resources developed for Turkish over the last two decades. (Parts of this chapter were previously published as Oflazer (Lang Resour Eval 48(4):639–653, 2014).)

1.1 Introduction

Turkish is a language in the Turkic family of Altaic languages which also includes Mongolic, Tungusic, Korean, and Japonic families. Modern Turkish is spoken mainly by about 60M people in Turkey, Middle East, and in Western European countries. Turkic languages comprising about 40 languages some of which are extinct are spoken as a native language by 165–200M people in a much wider geography, shown in Fig. 1.1. Table 1.1 shows the distribution of Turkic speakers to prominent members of the Turkic family.

Turkish and other languages in the Turkic family have certain features that pose interesting challenges for language processing. Turkish is usually used as a textbook example while discussing concepts such as agglutinating morphology or vowel harmony in morphophonology, or free constituent order in syntax. But there are many other issues that need to be addressed for robust handling language processing tasks.

K. Oflazer (✉)
Carnegie Mellon University Qatar, Doha-Education City, Qatar
e-mail: ko@cs.cmu.edu

M. Saraçlar
Boğaziçi University, Istanbul, Turkey
e-mail: murat.saraclar@boun.edu.tr

K. Oflazer, M. Saraçlar (eds.), *Turkish Natural Language Processing*,
Theory and Applications of Natural Language Processing,
https://doi.org/10.1007/978-3-319-90165-7_1

Fig. 1.1 The geography of Turkic languages (Source: Wikipedia), https://en.wikipedia.org/wiki/Turkic_languages, accessed 26 April 2018

Table 1.1 Distribution of speakers of Turkic languages (Data source: Wikipedia, https://en.wikipedia.org/wiki/Turkic_languages, accessed 26 April 2018)

Language	Percentage (%)
Turkish	30.3
Azerbaijani	11.7
Uzbek	10.2
Kazakh	4.3
Uyghur	3.6
Tatar	2.2
Turkmen	1.3
Kyrgyz	1.0
Other	35.4

Despite being the native language of over 60M speakers in a wide geography, Turkish has been a relative late-comer into natural language processing and development of tools and resources for Turkish natural language processing has only been attempted in the last two decades. Yet Turkish presents unique problems for almost all tasks in language processing ranging from tag-set design to statistical language modeling, syntactic modeling, and statistical machine translation, among many others. On the other hand, solutions to problems observed for Turkish when appropriately abstracted turn out to be applicable to a much wider set of languages. Over the years many tools and resources have been developed but many more challenges remain: For example, there are no natural sources of parallel texts where one side is Turkish (akin to say Europarl parallel corpora), so researchers working on statistical machine translation can only experiment with rather limited data which will not increase to the levels used for pairs such as English-Chinese or English-Arabic any time soon. Other more mundane issues such as drifting away from a one-to-one correspondence between orthography and pronunciation due to the recent wholesale import of words from other languages such as English with their native orthography *and* pronunciation, cause rather nasty problems even for the basic stages of lexical processing such as morphology. For example, one usually sees words like *serverlar* (servers) where, as written, the vowels violate the harmony constraints, but as pronounced, they don't, because of a bizarre assumption by the writers of such words that the readers will know the *English* pronunciation of the root words for the vowel harmony to go through!

Nevertheless, despite these difficulties the last several years have seen a significant increase of researchers and research groups who have dedicated efforts into building resources and addressing problems and the future should be quite bright moving forward.

In this introductory chapter we present a bird's eye view of relevant aspects of Turkish important from a language and speech processing perspective. Readers interested in Turkish grammar from more of a linguistics perspective may refer to, e.g., Göksel and Kerslake (2005).

1.2 Turkish Morphology

Morphologically Turkish is an agglutinative language with morphemes attaching to a root word like "beads-on-a-string." There are no prefixes and no productive compounding (e.g., as found in German) and most lexicalized compounds have non-compositional semantics (e.g., *acemborusu*, literally *Persian pipe*, actually is the name of a flower.)

Words are formed by very productive affixations of multiple suffixes to root words from a lexicon of about 30K root words excluding proper names. The noun roots do not have any classes nor are there any markings of grammatical gender in morphology and syntax. The content word root lexicons have been heavily influenced by Arabic, Persian, Greek, Armenian, French, Italian, German

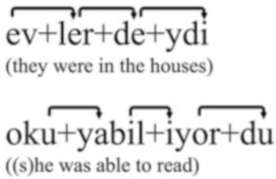

Fig. 1.2 Two examples of the cascaded operation of vowel harmony (Oflazer 2014) (Reprinted with permission)

and recently English, owing to the many factors such as geographical, cultural, commercial, and temporal proximity. Literally overnight, the alphabet used for writing the language was switched from the Arabic alphabet to a Latin alphabet in 1928, and this was followed by a systematic replacement of words of Arabic, Persian, and sometimes western origins, with native Turkish ones, but many such words still survive.

When used in context in a sentence, Turkish words can take many inflectional and derivational suffixes. It is quite common to construct words which correspond to almost a sentence in English:

yap+abil+ecek+se+k → if we will be able to do (it)

Almost all morphemes have systematic allomorphs that vary in respective vowels and sometimes in boundary consonants. For example, in

paket+ten (from the package) vs. *araba+dan* (from the car)

we see an example of a consonant assimilating at the morpheme boundaries and vowels in morphemes "harmonizing" with the previous vowel. Vowel harmony in fact operates from left-to-right in a cascaded fashion as shown in Fig. 1.2. Oflazer (1994) presents details of Turkish morphophonology as implemented in a two-level morphology setting (Koskenniemi 1983). Many relevant aspects of Turkish morphology will be covered in Chap. 2.

Multiple derivations in a given word are not an uncommon occurrence. Arısoy (2009) cites the word *ruhsatlandırılamamasındaki* as a word with nine morphemes, observed in a large corpus she worked with. The word roughly means *related to (something) not being able to acquire certification*, and is used as a modifier of some noun in context. Internal to the word, there are five derivations as shown in Fig. 1.3, where we start with a root word *ruhsat* (certification) and after five derivations end up as a modifier.

But in general things are saner: The average number of bound and unbound morphemes in a word in running text is about three but this is heavily skewed. Also, on the average, each word has about two different morphological interpretations due to root having multiple parts-of-speech, homography of some suffixes, and multiple segmentations of a given word into morphemes.

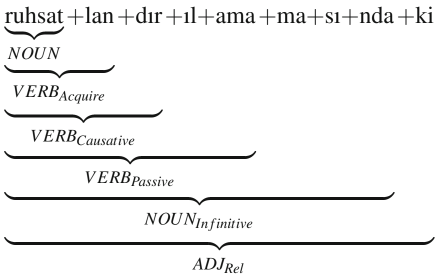

Fig. 1.3 Derivations in a complex Turkish word (Oflazer 2014) (Reprinted with permission)

Table 1.2 Morpheme count and morphological ambiguity in the most frequent 20 Turkish words (Oflazer 2014) (Reprinted with permission)

	Word	Morphemes	Ambiguity		Word	Morphemes	Ambiguity
1	bir	1	4	11	kadar	1	2
2	bu	1	2	12	ama	1	3
3	da	1	1	13	gibi	1	1
4	için	1	4	14	ol+an	2	1
5	de	1	2	15	var	1	2
6	çok	1	1	16	ne	1	2
7	ile	1	2	17	sonra	1	2
8	en	1	2	18	ise	1	2
9	daha	1	1	19	o	1	2
10	ol+arak	2	1	20	ilk	1	1

Table 1.2 shows the 20 most frequent words in a large Turkish corpus, along with the number of morphemes in the word and morphological ambiguity for each. We can estimate from these numbers that, since the more frequent words have just one morpheme, many of the lower frequency words have more than three or more morphemes. Also, most of the high-frequency words have relatively high morphological ambiguity, which, for words with one morpheme, corresponds to having different root parts-of-speech. Hence an average of two morphological interpretations mentioned above means that morphological ambiguity for words with many morphemes (owing usually to, for example, segmentation ambiguity) is actually less.

Another aspect of Turkish morphology is the heavy use for derivational morphemes in word formation as exemplified in Fig. 1.3. Table 1.3 shows the number of possible word forms (including inflected variants) that can be generated from only *one* noun or a verb root using zero, one, two, and three derivational morphemes, with the zero case counting only the basic inflectional variants. The total column shows the cumulative number of word forms with up to the number of derivations

Table 1.3 Number of words that can be derived using 0, 1, 2, or 3 derivational morphemes (Oflazer 2014) (Reprinted with permission)

Root	# derivations	# words	Total
masa	0	112	112
(Noun, (*table*))	1	4663	4775
	2	49,640	54,415
	3	493,975	548,390
oku	0	702	702
(Verb, (*read*))	1	11,366	12,068
	2	112,877	124,945
	3	1,336,266	1,461,211

on the same row.[1] It is certain that many of these derived words are never used but nevertheless, the generative capacity of the morphological processes can generate these. The fact that a given verb root can give rise to about 1.5M different word forms is rather amazing.[2] To tame this generative capacity, the derivational processes need to be semantically constrained which is extremely hard to do in a morphological analyzer.

Sak et al. (2011) present statistics from a large corpus of Turkish text of close to 500M Turkish words collected from mainly news text. They find about 4.1M unique words in this corpus, with the most frequent 50K/300K word forms covering 89%/97% of the words, respectively, and 3.4M word form appearing less than 10 times and 2M words appearing only once. The most crucial finding is that while increasing the corpus size from 490M to 491M by adding a text of 1M words, they report encountering 5539 new word forms not found in the first 490M words!

Figure 1.4 from Sak et al. (2011) shows the number of distinct stems and the number of distinct morpheme combinations that have been observed in this corpus. One can see that at around 360M words in the corpus, the number of distinct morpheme combination observed reaches around 46K *and* exceeds the number of distinct stems observed. This leads to an essentially infinite lexicon size and brings numerous challenges in many tasks.[3]

[1]These numbers were counted by using the *xfst*, the Xerox finite state tool (Beesley and Karttunen 2003), by filtering through composition by restricting output by the respective root words and with the number of symbols marking a derivational morpheme, and then counting the number of possible words.

[2]See Wickwire (1987) for an interesting take on this.

[3]It turns out that there are a couple of suffixes that can at least theoretically be used iteratively. The causative morpheme is one such morpheme, but in practice up to three could be used and even then it is hard to track who is doing what to whom.

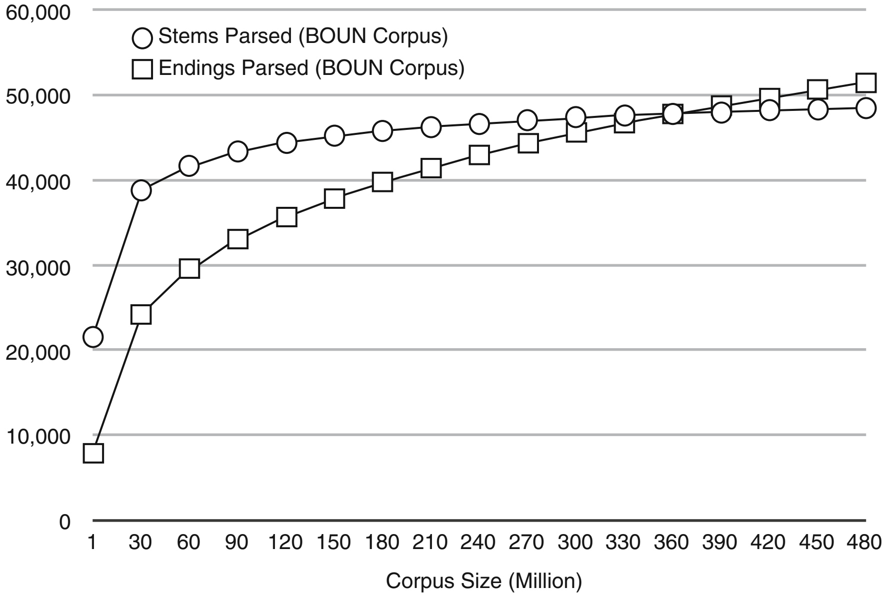

Fig. 1.4 Growth of number of unique stems and endings with corpus size (Sak et al. 2011) (Reprinted with permission)

1.3 Constituent Order and Morphology-Syntax Interface

The unmarked constituent order in Turkish is *Subject–Object–Verb* with adjuncts going in more or less freely anywhere. However all six constituent orders are possible with minimal constraints.[4] As is usual with other free constituent order languages, the freeness comes with the availability of case marking on the nominal arguments of the verbs.

The following are examples of constituent order variations along with the contextual assumptions when they are used. In all cases, the main event being mentioned is *Ekin saw çağla*, with the variations encoding the discourse context and assumptions.

- Ekin Çağla'yı gördü. (*Ekin saw Çağla.*)
- Çağla'yı Ekin gördü. (*It was Ekin who saw Çağla.*)
- Gördü Ekin Çağla'yı. (*Ekin saw Çağla (but was not really supposed to see her.)*)
- Gördü Çağla'yı Ekin. (*Ekin saw Çağla (and I was expecting that)*

[4]One constraint usually mentioned is that indefinite (and nominative marked) direct objects move with the verb, but there are valid violations of that observed in speech (Sarah Kennelly, personal communication).

- Ekin gördü Çağla'yı. (*It was Ekin who saw Çağla (but someone else could also have seen her.)*)
- Çağla'yı gördü Ekin. (*Ekin saw Çağla (but he could have seen someone else.)*)

Handling these variations in the usual CFG-based formalisms is possible (though not necessarily trivial or clean). Çetinoğlu's large scale LFG grammar for Turkish (Çetinoğlu 2009), developed in the context of the Pargram Project (Butt et al. 2002), handled these variations in a principled way but did not have a good way to encode the additional information provided by the constituent order variations.

A more interesting impact of complex morphology especially derivational morphology is on modeling syntactic relationships between the words. Before elaborating on this, let's describe an abstraction that has helped us to model these relationships.

The morphological analysis of a word can be represented as a sequence of tags corresponding to the morphemes. In our morphological analyzer output, the tag `^DB` denotes derivation boundaries. We call the set of morphological features encoded between two derivations (or before the first of after the last, if any) as an *inflectional group* (IG). We represent the morphological information in Turkish in the following general form:

$$\texttt{root+IG}_1 + \texttt{\^{}DB+IG}_2 + \texttt{\^{}DB+}\cdots + \texttt{\^{}DB+IG}_n.$$

where each $\texttt{IG}_i$ denotes the relevant sequence of inflectional features including the part-of-speech for the root (in IG_1) and for any of the derived forms.[5] A given word may have multiple such representations depending on any morphological ambiguity brought about by alternative segmentations of the word, and by ambiguous interpretations of morphemes.

For instance, the morphological analysis of the derived modifier `uzaklaştırılacak` ("(the one) that will be sent away," literally, "(the one) that will be made to be far,") would be:[6]

```
uzak+Adj
      ^DB+Verb+Become
      ^DB+Verb+Caus
      ^DB+Verb+Pass+Pos
      ^DB+Adj+FutPart+Pnon
```

[5] Although we have written out the root word explicitly here, whenever convenient we will assume that the root word is part of the first inflectional group.

[6] *uzak* is far/distant; the morphological features other than the obvious part-of-speech features are: `+Become`: become verb, `+Caus`: causative verb, `+Pass`: passive verb, `+Pos`: Positive Polarity, `+FutPart`: Derived future participle, `+Pnon`: no possessive agreement.

spor arabanızdaydı
sports car-your-in DB it-was

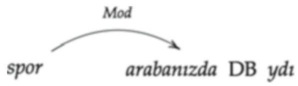

Fig. 1.5 Relation between inflectional groups

The five IGs in this word are:

1. `uzak+Adj`
2. `+Verb+Become`
3. `+Verb+Caus`
4. `+Verb+Pass+Pos`
5. `+Adj+FutPart+Pnon`

The first IG indicates that the root is a simple adjective. The second IG indicates a derivation into a verb whose semantics is "to become" the preceding adjective (equivalent to "to move away" in English). The third IG indicates that a causative verb (equivalent to "to send away" in English) is derived from the previous verb. The fourth IG indicates the derivation of a passive verb with positive polarity, from the previous verb. Finally, the last IG represents a derivation into future participle which will function as a modifier of a nominal in the sentence.

We can make two observations about IGs: (1) the syntactic relations are NOT between words, but rather between IGs of different words, and (2) the role of a given word in the sentence is determined by its last IG! To further motivate this, we present the example in Fig. 1.5. The second word in the phrase *spor arabanızdaydı* ("it was in your sports car") has a second/final IG which happens to have the part-of-speech of a verb. However there is also the adjective-noun construction *spor araba-* (sports car), where the word *spor* acts as a modifier of *araba*. So the modification relation is between (the last IG of) *spor* and the first IG of the next word (which has the part-of-speech noun) and not with the whole word whose final part-of-speech is a verb. In fact, different IGs of a word can be involved in multiple relations with different IGs of multiple words as depicted in a more comprehensively annotated sentence in Fig. 1.6.[7] In Fig. 1.6, the solid lines denote the words and the broken lines denote the IGs in the words. Note that in each case, a relation from a dependent emanates from the last IG of a word, but may land on any IG as the head. The morphological features encoded in the IGs are listed vertically under each IG with different IGs' features separated by vertical dashed lines. For instance, if we zoom into the three words in the middle of the sentence (shown in Fig. 1.7), we can note the following: The word *akıllısı* is composed of three IGs; it starts as noun *akıl* ("intelligence"), and with the derivational suffix *+li*, becomes an adjective ("with intelligence/intelligent") and then through a zero derivation becomes again a noun ("one who is intelligent"). The word *öğrencilerin* (of the students) and this final IG of *akıllısı* have the necessary morphological markings and agreement features to

[7]Here we show surface dependency relations, but going from the dependent to the head.

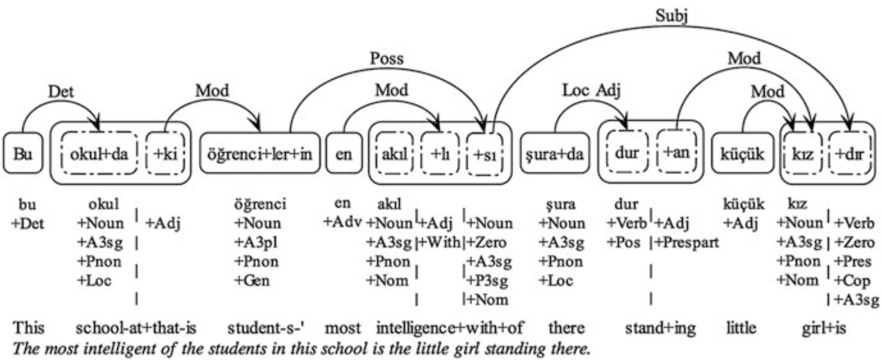

Fig. 1.6 Relations between IGs in a sentence (Oflazer 2014) (Reprinted with permission)

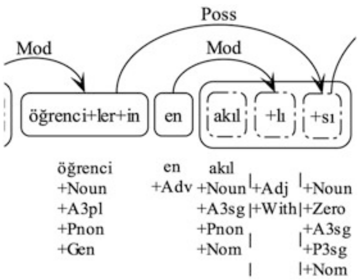

Fig. 1.7 Multiple syntactic relations for a word in Fig. 1.6 (Oflazer 2014) (Reprinted with permission)

form a possessor/possessee noun compound, and this is indicated by the relation by *Poss*. The more interesting example is the adverbial intensifier *en* ("most") modifying the intermediate IG with the part-of-speech adjective—it cannot have any other relationship, adverbials modify adjectives and not nouns. Thus we get a noun phrase meaning "the most intelligent of the students."

We have used IGs as a convenient abstraction in both statistical and rule-based contexts: Hakkani-Tür et al. (2002) modeled morphological disambiguation in terms of IGs. Çetinoğlu (2009) used IGs as basic units when modeling LFG syntax. Eryiğit et al. (2008) used IGs in the context of dependency parsing. The Turkish Treebank (Oflazer et al. 2003) has been encoded in terms of relations between IGs.

1.4 Applications

In this section we review some natural language and speech applications for Turkish, highlighting the challenges presented by Turkish in the context of these applications together with proposed solutions. While the applications span a wide spectrum, the challenges and solutions mostly follow a common theme. The complex morphology in combination with free word order and morphology–syntax interface summarized in the previous sections underlie the challenges. The solutions make use of morphological and morphosyntactic analysis to alleviate the challenges.

Spelling Checking and Correction Methods that rely on a finite list of words or a list of root words with some fixed number of affixes cannot capture lexicon of Turkish. We have developed efficient spelling correction algorithms for languages like Turkish based on error tolerant finite state recognition, operating on a finite state recognizer model of morphology that can encode an infinite number of words (Oflazer 1996).

Tagset Design It is not possible to fully represent the morphosyntactic information encoded in morphology with a finite set of tags. The data in Fig. 1.4 already hints at this. There are of course a small number of root part-of-speech categories but with multiple inflectional and derivational affixes affixed, the word may end up having many morphological features including multiple parts-of-speech, all of which may have syntactic implications. See Hakkani-Tür et al. (2002) for statistics on the number of different possible tags.

Syntactic Modeling As we saw in the previous section, derivational morphemes have interesting implications in syntactic modeling using either constituency based formalisms or dependency based formalisms. These will be discussed in more detail in Chaps. 7 and 9.

Statistical Language Modeling A large vocabulary size almost always leads to a data sparseness problem in word-based language modeling. This is especially important when the text corpora used for language model estimation are not extremely large. One approach to limit the vocabulary size and hence combat data sparseness is to use sub-lexical units instead of words in language modeling. Traditional n-gram language models predict the next unit given the history consisting of $n-1$ units. There is a trade-off between the length and the predictive power of the units used for traditional n-gram language models. On the one hand, the shorter the units, the more common they are. So data sparseness is less of an issue for shorter units. On the other hand, for shorter units, the history needs to include many more units for the same level of predictive power. This is easy to see when one compares letter-based language models with word-based language models.

Arısoy (2009) and Sak (2011) have investigated using sub-lexical units in language modeling for Turkish. Both morphological analyzers and unsupervised statistical word segmentation techniques yield sub-lexical units that improve the coverage and performance of statistical language models.

Although morphological analysis provides meaningful units useful for language modeling, it also has some issues. First, building a wide coverage morphological analyzer is costly and requires expert knowledge. Second, the coverage of the morphological analyzer is limited by the root lexicon and this is especially important for proper nouns. Finally, when using morphological analysis to obtain the sub-lexical language modeling units, an important issue is morphological disambiguation. For statistical language modeling, consistency in disambiguation can be as important as accuracy.

On the other hand, unsupervised word segmentation techniques typically require only a word list to come up with the sub-lexical language modeling units. However, these units do not necessarily correspond to actual morphemes and may not be as meaningful and informative as those obtained by morphological analysis. Further unsupervised statistical processing such as clustering can provide a way of improving the predictive power of these units.

In addition to the traditional language models that predict the next unit given the units in the history, feature based language models allow easy integration of other information sources. For Turkish, Arısoy (2009) incorporated morphological and syntactic features in language modeling both for lexical and sub-lexical units.

Details of these approaches will be covered in Chap. 4.

Pronunciation Modeling Applications that aim a conversion between text and speech require a way of determining how words are pronounced. For limited vocabulary applications, a hand-crafted pronunciation lexicon that simply lists the pronunciations of the words in the vocabulary is adequate. However, for Turkish, the large vocabulary size implies that a list of pronunciations for use in speech applications is rather inadequate.

Oflazer and Inkelas (2006) describe a computational pronunciation lexicon capable of determining the position of the primary stress. Their implementation uses a series of finite state transducers including those for two level morphological analysis, grapheme-to-phoneme mapping, syllabification, and stress computation. They also report that for a corpus of about 11.6 million tokens, while the average distinct morphological parses per token is 1.84, the average distinct pronunciations per token is 1.11 when taking stress into account and only 1.02 ignoring stress. The implications of this analysis for speech applications will be discussed below.

Automatic Speech Recognition (ASR) In addition to the challenges related to statistical language modeling, ASR (or STT) systems also have to deal with issues related to pronunciation modeling. In particular, the mainstream ASR systems make use of phone-based acoustic models that require a pronunciation lexicon to map words into phone sequences. While information about the position of stress can improve the acoustic models, the use of stress is not vital and common for ASR.

As mentioned above, while a pronunciation lexicon can be built by hand for medium vocabulary sizes, large vocabulary continuous speech recognition (LVCSR) requires an automatic process for building the pronunciation lexicon such as the one implemented by Oflazer and Inkelas (2006). Although the process of mapping the graphemic representation to the phonetic representation is not overly complicated, it does require morphological analysis. Their observation that over 98% of the tokens have a single pronunciation when the position of the primary stress is ignored, and that the remaining tokens have only two alternative pronunciations (differing mostly in vowel length and consonant palatality), suggests that pronunciation disambiguation is not really necessary for ASR.

An alternative approach uses grapheme-based acoustic models and lets the context-dependent graphemic acoustic models implicitly take care of pronunciation modeling. While graphemic acoustic modeling might seem somewhat simplistic, it works quite well in practice for languages where the orthography is not far from the pronunciations.

Using a sub-lexical language model further complicates pronunciation modeling. When morphological analysis is used to obtain the sub-lexical units, it is not possible to determine the pronunciation of a sub-lexical item without looking at its context, the vowels of most suffixes are determined by vowel harmony and adding a suffix may change the pronunciation of the root. Therefore, the pronunciation lexicon will have to include multiple pronunciations complicating the system and allowing for incorrect pronunciations. This issue is even more dramatic when unsupervised word segmentation is used to obtain the sub-lexical units. Some units may not even be pronounced. As graphemic acoustic models do not require a phonetic representation, no further complications arise from using sub-lexical units for language modeling.

Acoustic confusability is another issue that needs to be considered when using sub-lexical units. Longer units are less confusable than shorter units simply because their acoustic neighborhood is less populated. Acoustic confusability and the trade-off for language modeling discussed above suggest that short units are not preferable for ASR.

For the Turkish broadcast news transcription task (Saraçlar 2012), using context-dependent grapheme-based acoustic models, and a language model based on a vocabulary of 76K sub-lexical units with an average unit length of 7–8 letters gives a very good coverage and the lowest word error rate percentage (Arısoy et al. 2009).

Speech Retrieval Speech retrieval systems combine ASR with information retrieval (IR). The IR component typically forms an index from the output of the ASR system and searches this index given the user query. While obtaining a simple text output from the ASR system makes it possible to directly leverage text retrieval techniques, using alternative speech recognition hypotheses in the form of a lattice has been shown to significantly improve retrieval performance (Chelba et al. 2008).

For Turkish, Arısoy et al. (2009) investigated spoken term detection (or keyword search) for Turkish broadcast news. Parlak and Saraçlar (2012) further extended this work and also built a spoken document retrieval system for the same task.

Since queries tend to include rare words, the frequency of queries containing words that are outside the vocabulary of the ASR system can be quite high, especially for Turkish. In order to deal with these queries it is common to make use of sub-lexical units even when the ASR system produces word-based outputs. Of course, the same sub-lexical units used for ASR can also be used for indexing and search. Arısoy et al. (2009) have shown that the best performance for Turkish broadcast news retrieval is obtained by combining the output of systems based on word and sub-word units.

Another common technique utilized especially for spoken document retrieval is stemming. While it is possible to determine the stem using full morphological analysis, stemming is actually an easier task. For both text and speech document retrieval using the first five characters of a word was shown to perform well (Can et al. 2008; Parlak and Saraçlar 2012).

Speech Synthesis or Text-to-Speech (TTS) Text-to-Speech systems require a text analysis step in order to obtain a phonetic representation enriched with stress and prosodic markers for a given input text. Determining the pronunciation of a word sequence together with the required stress and prosodic information is more involved than building a pronunciation lexicon for ASR.

Oflazer and Inkelas (2006) report that, when taking the primary stress into account, about 90% of the tokens have a single pronunciation, about 9% have two distinct pronunciations and the rest have three or more pronunciations. Therefore, pronunciation disambiguation is a required component for the text analysis component of a TTS system. Külekçi (2006) analyzed the pronunciation ambiguities in Turkish and suggested that morphological disambiguation (MD), word sense disambiguation (WSD), and named entity recognition (NER) can be used for pronunciation disambiguation.

Statistical Machine Translation Just as with statistical language modeling, a large vocabulary implies sparseness in statistical machine translation, which is compounded by the fact that no really large parallel corpora involving Turkish exist to offset this. Thus approaches exploiting morphology in various ways have been proposed with good improvements over word-based baseline.

At this point, it should be clear that morphology is bound to create problems for three components of a statistical machine translation systems for Turkish. Let's look at a rather contorted but not that unreasonable example of a hypothetical process of how an English phrase becomes a Turkish word in the ideal case. Figure 1.8 shows how different parts of the English phrase (mostly function words) are scrambled around and then translated into morphemes which when concatenated gives us a single word *sağlamlaştırabileceksek*. One can immediately see that the process of alignment—the starting point for training SMT systems—is bound to

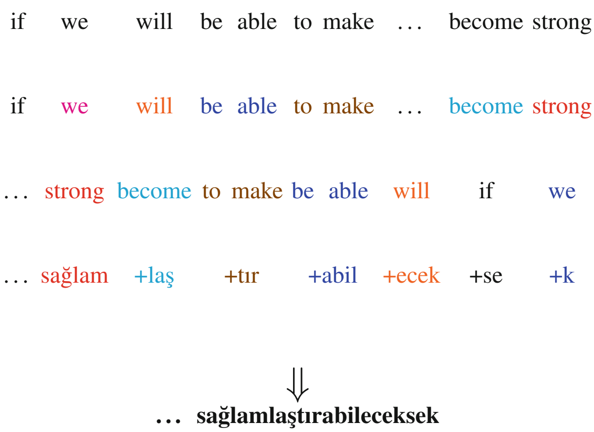

Fig. 1.8 How English becomes Turkish in translation (Oflazer 2014) (Reprinted with permission)