Spoken Turkish Corpus Project

Yuji KAWAGUCHI

Tokyo University of Foreign Studies

Bosphorus University, Istanbul Oct. 4 2012

Presentation flow

1. Corpus Linguistics
2. Turkish Corpus
3. Constructing spoken corpus
4. Spoken Turkish Corpus
5. Corpus-based analysis

Concluding remark

1. Corpus Linguistics

1967 Brown Corpus
Computational Analysis of
Present-Day American English
1987 Collins' COBUILD dictionary
1991 British National Corpus
American National Corpus

1985 A Comprehensive Grammar of the
English Language
1999 Longman Grammar of Spoken and
Written English

2.1. Turkish Corpus

- Written Turkish Corpus:
  METU Turkish Corpus
    http://fodor.ii.metu.edu.tr/content/metu-turkish-corpus
  Turkish National Corpus
    http://www.tnc.org.tr/index.php

- Spoken Turkish Corpus:
  Spoken Turkish Corpus (STC)
    http://std.metu.edu.tr/en

2.2. COE Projects at TUFS

21 Century COE Program, Usage-Based Linguistic Informatics (UBLI) 2002-2006

Global COE Program, Corpus-based Linguistics and Language Education (CbLLE) 2007-2011

Construction of Spoken Language Corpora

-French, with Aix-Marseille University

-Spanish, Autonomous University of Madrid

-Malay, Kebangsaan University

-Chinese, Tamkang University

-Russian, Russian Satate University for Humanities

http://www.coelang.tufs.ac.jp/english/index.html

2.3. Spoken Turkish Corpus Project

21st Century COE Program
Selim Yılmaz (Marmara University),
Arsun-Uras Yılmaz
(University of Istanbul)
2005: 29 free conversations 98,100 tokens

Global COE Program
2007: 14 free conversations 56,300 tokens
2008: 10 free conversations 38,900 tokens
2009: 23 free conversations 75,900 tokens
2010: 30 free conversations 60,500 tokens
2011: 59 free conversations 184,700 tokens
Total: 514,400 tokens

http://cblle.tufs.ac.jp/multilingual_corpus/tr/index.html?contents_xml=top&menulang=en

3. Constructing spoken corpus

Objectives

- Written language vs Spoken language
  → Syntactic/morphogical/phonological structures
- Description of real use of language
  → Sociolinguistic/pragmatic variations
- Usage norm for language teaching
  → Frequent/infrequent use

3.1. Nature of spoken language

- Does “spoken” mean “broken”?
  → Claire Blanche-Benveniste
- What is the unit of speech?
  → Sentence, discours, turn
- Enormous variations
  → Who speaks to whom and where?
- Representativeness
  → Is there a representative spoken speech?

3.2. Informant's profile

*No	*Date		
Name / Family Name		Sex	☐ Man ☐ Woman
Address			
Telephon	e-Mail		
Profession	(Institution)	Birth Place (City/Country)	
Birth Date	Education	☐Primary School ☐Middle School ☐High School ☐University ☐ Others (.....)	
1st Language	☐Turkish ☐ Other languages (.....)		
If you have another language in everyday life:			
Parents' Birth Place Father: (City / Country) Mother: (City / Country)			
*No and subject of conversation :			
*Particular information:			

3.3. Transcription of spoken language

Symbol	Meaning
Initials of Name	Ex: SY = Selim Yılmaz. SY1, SY2, SY3
(?), (!)	Interrogation and Exclamation
-	Beginning of discours. Ex: -SY1....
#, ##, ###...	Pause. Symbol represents its length.
ooo, aaa,...	Long and emotional vowel Ex: yook, haayır, eveet, yaaa.
(.)	Incompletion Ex: çek(.) çekilmez gibi geliyo(r)

3.3. Transcription of spoken language

(x) (xy)	Dropping Ex: bur(a)da, bi(r) gün, geliyo(r)
_____	Accentuation Ex:## ancak Kastamonu'nun çevresi ilçeleri çok güzel
e, ee, eee,…	Hesitation or filler. Ex: yoksa ee yani çek(.) çekilmez gibi geliyo(r)
m, mm,…	Murmuring
{…..}	Dislocated element. Ex: hava güzel {bugün}, yarın gideceğim {okula}.
[…..]	Paralinguistic element or gesture Ex: [gülüşmeler], [öksürme], [gürültü]

3.3. Transcription of spoken language

<.....>	Overlap. EX: -ben de yani <ya öyle> biraz yüzeysel düşünmüşüz ama içine girince <tabi çok> çok derin olduğunu fark ettik(.)
.....	Accelerated element EX: *hakikaten de güzel birşeyler yaptık*
(...)	Incomprehensive element EX: doktora yapıyordu internetten indiriyor (...) yapıyor
%.....%	Special intonation EX: çünki el yazmaları var ## %bunun dışında% doğa güzelliği mükemmel {Kastamonu'da}

4. Spoken Turkish Corpus

Example 1 : Lecture on Turkish Literature

ND1 – “geç kaldım # bütün bunlar için geç kaldım” # “yoo böyle deme gençsin # öğrenmenin yaşı yoktur # bittiği durduğu yer yoktur” # bir gün konuşma içerisinde geçti (.) arasında *böyle ilmi yetersizligini görürüz* bur(a)da bir bozkırın ötesi diyor # yani öyle bi(r) (...)

OS1 - evet biraz önce konustuğumuz gibi # yani ufkunu biraz daha ötelere götürmeye çalışıyor # yani ama demek ki onun ilmide o kadar yeterli değil # yani kendiside bunun farkında #

ND2 - bozkırın ötesinde (.) yerleşik hayat şimdi Nizam Bey bozkırda yetişmiş çadırda yaşadım falan diyor (.) daha yerleşik hayat istiyor {yani}

.....

4. Spoken Turkish Corpus Example 2 : About the subway

FS19 - yani mimar olarak ben şöyle düşünebiliyorum e tamam şimdi metrobüsler gayet güzel yer altına yapılan bu sistem çok pahalı bunu onaylamıyorum yer altına gerek yok bunun yerine artık tüm dünyada kabul edilmiş artık hep yer üstünde ee asma olarak yapılması lazım hem pratik hem ekonomik ama bunun yerine bizde nedense hep altlardan kazmayı düşünüyorlar daha zorunu yapmaya çalışıyorlar bu heralde siyasetin bi(r) parçası veya başka bi(r) insanlara imtiyaz tanımanın bi(r) parçası gibi geliyor {bana}

……

5. Corpus-based analysis

5.1. Dropping of /r/

Kawaguchi (2009)

Corpus 2005-2006

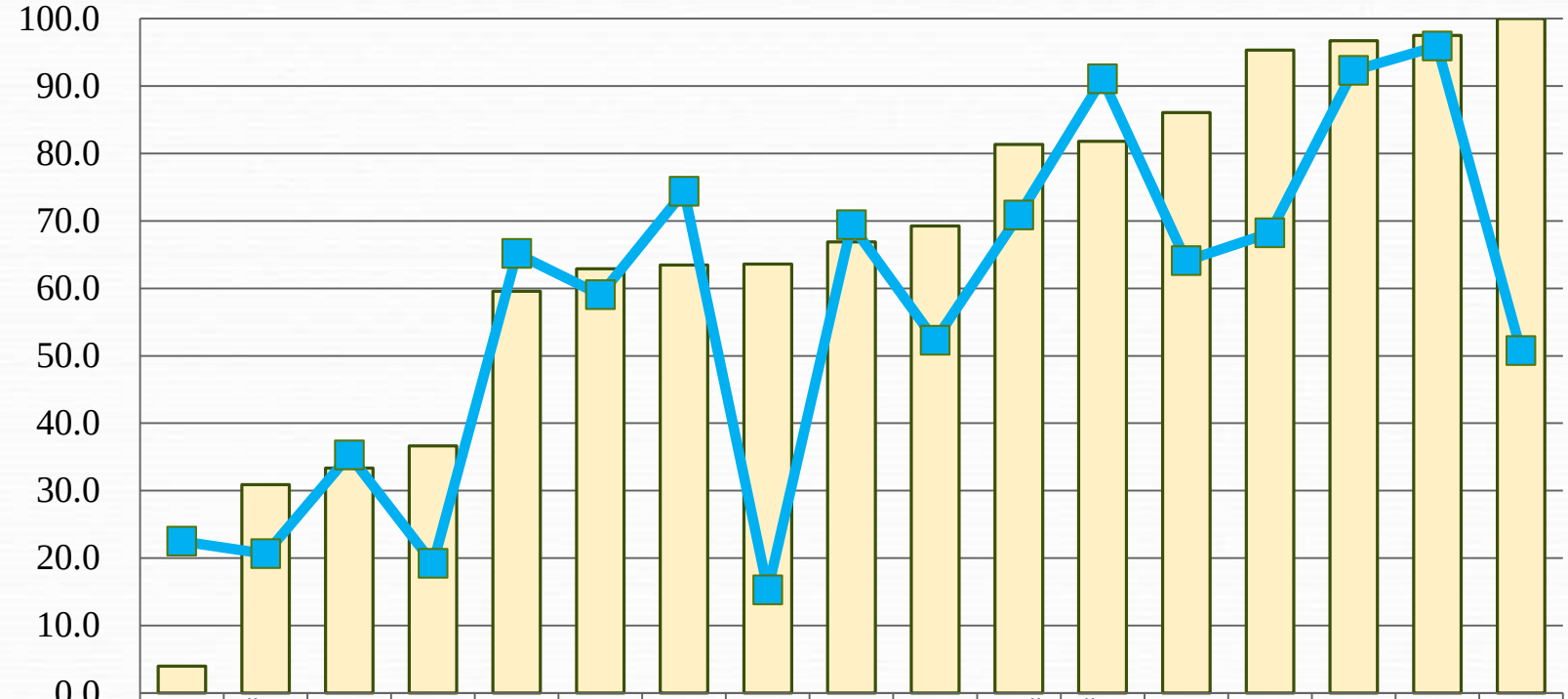
10h26m, word types 17,417, word tokens 93,837

Different parameters

- Conversation topic : r = .11
- Conversation speed : r = .25
- Gender difference : r = .16
- Age difference : r = - .19

(r = Pearson correlation coefficient)

Individual variation of -r dropping in *-Iyor* and *bi(r)*



	BG	ÜD	FK	ŞK	HS	RŞŞ	MG	SY	AHT	EY	MÜ	ÖK	SG	AK	BH	DH	NS
yor	4.0	30.9	33.3	36.6	59.6	62.9	63.5	63.6	66.9	69.2	81.4	81.8	86.1	95.3	96.7	97.5	100.
bi(r)	22.5	20.6	35.3	19.2	65.1	59.0	74.4	15.3	69.4	52.3	70.8	91.0	64.1	68.2	92.3	95.9	50.7

|----------| **less –r dropped**

average of –r dropping: -Iyor = 69.0 %; bir = 61.8 %

5.2. Postpredicative element

Kawaguchi (2011)

Oya	bir köpek	gördü	bahçede.
Utterance Head Theme	Immediately before Predicate Focus	Predicate	Postpredicate Background

Syntax and Information Structure

5.2. Postpredicative element

	token
1. adverb	149
2. conjunctive	68
3. locative	61
4. dative	47
5. definite direct object	45
6. subject pronoun	40
7. noun phrase	37

	token
8. postposition	34
9. appelative	22
10. instrumental	19
11. genitive	17
12. interjection	14
13. verbal noun	7
14. others	6
total	566

5.2. Postpredicative element

	token	type
1.truth	41	açıkçası (11), aslında (17), esasında (3), gerçekten (5), hakikaten (2), etc.
2.time	25	birbirinden (1), -ken (2), sonra (2), şimdi (2), zaman (6), zamanda (2), etc.
3. opinion	14	bence (12), kanımca (1), sence (1)
4. example	14	mesela (14)
5. quantity	12	biraz (7)
6. place	1	aşağı (1)
7. sentential	42	artık (5), belki (5), herhalde (5), sanki (3), zaten (8)
Total	149	( ) occurrences

1) HS142- (...) # ben o yüzden buraya indim # {**açıkçası**}
AA20- (...) # kolay değil {**aslında**}

3) ES27- (...) ## bunları anlamak gerçekten hani ee ## bizim boynumuzun borcu {**bence**}
BY1 – (...) ee bunlara bunlara bi(r) düzen getirilmesi gerekiyo(r) {**kanımca**} #

Concluding remark

Corpus-based analysis of Spoken Turkish Corpus

- Structural analysis
  Phonological, morphological, syntactic variations
- Pragmatics & conversation analysis
  Fillers, discourse analysis, politeness
- Sociolinguistics & stylistics
  Diastratic & diaphasic variations
- Applied linguistics
  Norm of usage & authentic material for learners

References

Blanche-Benveniste, Claire. 1997. *Approches de la langue parlée en français,* Gap: Ophrys.

Erguvanlı, T. Eser. (ed.) 2001. *The Verb in Turkish,* Amsterdam/Philadelphia: John Benjamins.

Gibbon, Dafydd, Roger Moore and Richard Winski (eds.) (1998) *Handbook of Standards and Resources for Spoken Language Systems*, Vol.1-4, Berlin/New York: Mouton de Gruyter.

Göksel Aslı and Celia Kerslake . 2005. *Turkish A Comprehensive Grammar,* London: Routlege.

Göksel, Aslı and Özsoy Sumru A. (2000) "Is there a focus position in Turkish?", In *Studies on Turkish and Turkic Languages,* Aslı Göksel and Celia Kerslake (eds.) Wiesbaden: Harrasotiwz Verlag. 219-228.

Kawaguchi, Yuji. 2004. "Two Turkish Clause Linkages: *-DIK-* and *–mE-* -A pilot analysis based on the METU Turkish Corpus-" In: Toshihiro Takagaki. Susumu Zaima. Yoichiro Tsuruga. Francisco Moreno-Fernández and Yuji Kawaguchi (eds.) 2005. *Corpus-Based Approaches to Sentence Structures.* Amsterdam: John Benjamins, 151-177.

Kawaguchi, Yuji. 2009. "A Corpus-Driven Analysis of –r Dropping in Spoken Turkish", Y. Kawaguchi, M. Minegishi, J. Durand, (eds.) *Corpus Analysis and Variation in Linguistics*, Amsterdam/Philadelphia: John Benjamins, 281-297.

Kawaguchi, Yuji. 2011. "Postpredicate Elements in Spoken Turkish: A Copus-based Analysis" (in Japanese), *Working Papers in Corpus-based Linguistics and Language Education* 7, 171-197.

Morel, M.-A. et L. Danon-Boileau, 1998. *Grammaire de l'intonation. L'exemple du français*, Paris: Ophrys.

Özsoy, A. Sumru. 1999: *Türkçe Turkish,* ,İstanbul: Boğaziçi Üniversitesi Yayınları.

Özsoy, A. Sumru. 2004. *Türkçenin yapısı – I, Sesbilim,* İstanbul: Boğaziçi Üniversitesi Yayınları.

Yılmaz, Selim (2012) "Türkçede Sözlü Derlem Oluşturma Çalışmaları Üzerine Değerlendirmeler (Uluslararası Global COE Program Projesi Çerçevesinde)", *Working Papers in Corpus-based Linguistics and Language Education*, 9, Graduate School of Global Studies, Tokyo University of Foreign Studies, pp.165-184.
(http://cblle.tufs.ac.jp/assets/files/publications/working_ papers_09/section/165-184.pdf)

Thank you for your attention!

You can use our corpus exclusively for the academic purpose.

Please contact the COE office by e-mail :

cblle-faq@tufs.ac.jp

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers 24320102
Grant-in-Aid for Scientific Research(B) Person in charge:Yuji KAWAGUCHI.