

Traitement de corpus en turc parlé

Yuji KAWAGUCHI

Tokyo University of Foreign Studies

Selim YILMAZ

Université de Marmara

L'INALCO Paris le 10 décembre 2014

Plan

1. Linguistique du corpus
2. Corpus du turc
3. Corpus de la langue parlé
4. Corpus du turc parlé
5. Analyse basée sur le corpus

Conclusion

I. Linguistique de corpus

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

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

2.1. Corpus du turc

- Corpus du turc écrit:
METU Turkish Corpus

[http://fodor.ii.metu.edu.tr/content/
metu-turkish-corpus](http://fodor.ii.metu.edu.tr/content/metu-turkish-corpus)

Turkish National Corpus

<http://www.tnc.org.tr/index.php>

The screenshot shows the homepage of the Turkish National Corpus (TNC). At the top, it displays "Türkçe Ulusal Derlemi" and "Turkish National Corpus". Below this, there's a navigation bar with links for "About TNC", "WELCOME", "NEWS", "EVENTS", and "PAPERS". The "About TNC" section contains links for "Objective", "Content", "Annotation", "TNC Interface", "TNC (Demo Version)", "Publications", "Papers", "Theses", and "About Us". The "WELCOME" section says "Welcome to the homepage for Turkish National Corpus." and lists "13. Yıldırım Aksan, Mustafa Aksan, et al. (2012). Construction of the Turkish National Corpus (TNC). In Proceedings of LREC 2012, Istanbul, Türkiye." The "NEWS" section has a link to "Previous News". The "EVENTS" section lists "LREC 2012", "ICTL 2012", and "ITLT 2012". The "PAPERS" section lists "13. Yıldırım Aksan, Mustafa Aksan, et al. (2012). Construction of the Turkish National Corpus (TNC). In Proceedings of LREC 2012, Istanbul, Türkiye.". On the right side, there are links for "English", "TUD", "Homepage", "Links", "CONTACT", and "Previous Papers".

- Corpus du turc parlé:
Spoken Turkish Corpus (STC)
<http://std.metu.edu.tr/en>

The screenshot shows the homepage of the Spoken Turkish Corpus (STC). At the top, it displays the logo "STC Sözlü Türkçe Derlemi Spoken Turkish Corpus". Below this, there's a navigation bar with links for "Anasayfa", "Haberler", "Hakkında", "S.S.S.", "Gönüllü Bayevrusu", "Proje Ekibi", "Yayınlar", "Kaynakça", "EXMARaLDA", "Bağlantılar", "Teşekkürler", and "İndir". The main content area features a banner with the text "Gozden Geçirilmiş Sözlü Türkçe Derlemi Tanıtım Sürümü Kullanma Hazır!". It also includes sections for "EXMARaLDA 1-5", "Gönüllüler aranıyor!", and "Derlem nedir?". On the right side, there are language selection buttons for "Türkçe" and "English", a search bar, and a "Gönüllü Bayevru Formu".

2.2. Projets COE à TUFS

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

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

Construire les corpus de langues parlées
-Français, avec Université Aix-Marseille
-Espagnol, Université Autonome de Madrid
-Malais, Université Kebangsaan
-Chinois, Université Tamkang de Taiwan
-Russe, Université d'État des sciences humaines de Russie

The screenshot shows a website for the 21st Century COE Program "Usage-Based Linguistic Informatics" (2002–2006). The header features the project name in large letters, followed by a sub-header "Usage-Based Linguistic Informatics". Below this is a brief description of the project's aim: "The aim of this project is to innovate foreign language education by developing superior foreign language educational material and transmitting it through the Internet. Thus, an overall integration of theoretical and applied linguistics will be realized through the theoretical analysis of linguistic usages of different languages." On the left, there is a sidebar with a navigation menu including "Outline", "Projects", "Conference", "TUFS Language Modules", "TUFS e-learning system", "Multilingual Corpora", "Publications", and "Nurturing of Young Researchers". At the bottom left, there is a PDF file icon labeled "PDF File" and "COE Pamphlet" with links to "2004.12 >>" and "2005.10 >>". The right side of the page has a decorative background featuring a globe and various linguistic symbols.

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

2.3. Corpus du turc parlé

21st Century COE Program

Selim Yilmaz (Université de Marmara),
Arsun-Uras Yilmaz (Université d'Istanbul)

2005: 29 conversations : 98,100 mots

Global COE Program

2007: 14 conversations : 56,300 mots

2008: 10 conversations : 38,900 mots

2009: 23 conversations : 75,900 mots

2010: 30 conversations : 60,500 mots

2011: 59 conversations : 184,700 mots

2012: 17 conversations : 63,400 mots

Total: 577,800 mots

The screenshot shows a website for the "Turkish Multilingual Spoken Corpus". The header features a banner with a cityscape and the text "Turkish Multilingual Spoken Corpus". A sidebar on the left contains links: Top, Outline, Research Team, References, Corpus (which is highlighted in blue), Related Sites, and Pictures. The main content area has a "Top" button at the top right. Below it is a "Purpose" section with the following points:

- Elaboration of a large corpus of Spoken Turkish
- Transcription system for describing the characteristics of Spoken Turkish
- Data for several linguistic analyses (phonetic, syntactico-semantic and pragmatic)

Further down is a "About the quotation of the content of the description" section containing the following text:

The present corpus is published as one of the accomplishments of the 21st Century Center of Excellence (COE) Program "Usage-Based Linguistic Informatics" based at the Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies. A part of our multilingual corpus is shown below.) In case you quote the present corpus in your article, please remark the following references.

The screenshot shows a website for the "Global COE Multilingual Spoken Corpus". The header features a banner with the text "Global COE Multilingual Spoken Corpus" and "Corpus of spoken Turkish". A sidebar on the left contains links: Top, Outline 2005, Outline 2007, Research Team, References, Corpus 2005 (which is highlighted in blue), Corpus 2007, Related Sites, and Pictures. The main content area has a "Top" button at the top right. Below it is a "Purpose" section with the following points:

- Elaboration of a large corpus of Spoken Turkish
- Transcription system for describing the characteristics of Spoken Turkish
- Data for several linguistic analyses (phonetic, syntactico-semantic and pragmatic)

Further down is a "About the quotation of the content of the description" section containing the following text:

The present corpus is published as one of the accomplishments of two research projects based at the Institute of Global Studies, Tokyo University of Foreign Studies. The 21st COE Program "Usage-Based Linguistic Informatics" 2002-2006 and The Global COE Program "Corpus-based Linguistics and Language Education". A part of our multilingual corpus is presented below. In case you quote the present corpus in your article, please remark the following references.

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

3. Corpus de la langue parlée

Objectifs

- langue écrite vs langue parlée
 - structures syntaxique/morphologique/phonologique propres
- description de l'usage réel
 - variations sociolinguistique/pragmatique
- norme pour l'enseignement de langues
 - fréquence/erreur/acceptabilité

3.1. Nature de la langue parlée

- “parlé” veut dire “mauvais” ou “maladroite” ?
→ Claire Blanche-Benveniste
- quelle unité pour la langue parlée ?
→ phrase, discours, tour de parole
- énormes variations
→ qui parle à qui et où ?
- représentativité
→ langue parlée représentative, ça existe?

3.2. Profil d'informateur

*No	*Date		
Name / Family Name		Sex	<input type="checkbox"/> Man <input type="checkbox"/> Woman
Address			
Telephon	e-Mail		
Profession	(Institution)	Birth Place (City/Country)	
Birth Date	Education	<input type="checkbox"/> Primary School <input type="checkbox"/> Middle School <input type="checkbox"/> High School <input type="checkbox"/> University <input type="checkbox"/> Others (.....)	
1st Language	<input type="checkbox"/> Turkish <input type="checkbox"/> 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

Signe	Sens
Nom	Ex: SY = Selim Yılmaz. SY ₁ , SY ₂ , SY ₃
(?), (!)	Interrogation et exclamation
-	début du discours. Ex: -SY ₁
#, ##, ###...	pause, pause longue
ooo, aaa,...	voyelle émotionnelle longue Ex: yook, haayır, eveet, yaaa.
(.)	interruption Ex: çek(.) çekilmez gibi geliyor

3.3. Transcription

(x) (xy)	chute Ex: bur(a)da, bi(r) gün, geliyo(r)
_____	accentuation Ex:## <u>ancak</u> Kastamonu'nun çevresi ilçeleri çok güzel
e, ee, eee,...	hésitation Ex: yoksa ee yani çek(.) çekilmmez gibi geliyo(r)
m, mm,...	murmure
{.....}	dislocation Ex: hava güzel {bugün}, yarın gideceğim {okula}.
[.....]	éléments paralinguistiques Ex: [gülüşmeler], [öksürme], [gürültü]

3.3. Transcription

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

4. Corpus du turc parlé

Example 1 : Sur la littérature turque

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 yetersizliğini görürüz* bur(a)da bir bozkırın ötesi diyor # yani öyle bi(r) (...)

OS1 - evet biraz önce konustuğumuz gibi # yani ufkunu biraz daha otelere götürmeye çalışıyor # yani ama demek ki onun ilmi de o kadar yeterli değil # yani kendisi de bunun farkında #

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

.....
(2007-05)



Global COE Multilingual Spoken Corpus

Top
Outline2005
Outline2007
Research Team
References
Corpus2005
Corpus2007
Related Sites
Pictures

Global COE Program
Corpus-based Linguistics and Language Education

The 21st Century COE Program
Usage-Based Linguistic Informatics

Lecture on Turkish Literature

Turkish; “geç kaldım # bütün bunlar için geç kaldım” # “yoo böyle deme gençsin # öğrenmenin yaşı yoktur # bittiği durduğu yer yoktur” #

Turkish; bir gün konuşma içerisinde geçti (.) arasında *böyle ilmi yetersizliğini görürüz* bur(a)da bir bozkırın ötesi diyor # yani öyle bi(r)(...)

Turkish
 Japanese
 English

Scroll On/Off

4. Corpus du turc parlé

Example 2: Construction

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 her(h)alde siyasetin bi(r) parçası veya başka bi(r) insanlara imtiyaz tanımanın bi(r) parçası gibi geliyor {bana}

.....



Global COE Multilingual Spoken Corpus

Top
Outline2005
Outline2007
Research Team
References
Corpus2005
Corpus2007
Related Sites
Pictures

Global COE Program
Corpus-based Linguistics and Language Education

The 21st Century COE Program
Usage-Based Linguistic Informatics

Global COE Multilingual Spoken Corpus

About the Subway

Turkish; yani mimar olarak ben şöyle düşünebiliyorum e tamam şimdi metrobüsler gayet güzel yer altına yapılan sistem çok pahalı bunu onaylamıyorum yer altına gerek yok

Turkish; 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

Turkish
 Japanese
 English

Scroll On/Off

5. Analyse basée sur le corpus

5.1. Structure de l'énoncé oral turc

MARQUES DISCURSIVES (Déictique / Embrayeur)	VALEURS / FONCTIONS ENONCIATIVES
YANI	Explication / Explication / Paranthese (Continuité du discours)
EVET (BİRAZ ÖNCE) TAMAM (ŞİMDİ)	Confirmation (des propos de soi-même ou de l'autre) / Maintien de la parole
AMA (DEMEK Kİ)	Jugement supplémentaire à valeur contrastive / Antithèse / Possibilité de discordance
BEN / BANA	Point de vue du sujet parlant (jugement personnel) / Assertion / Prise en charge avec position égocentré

5.2. Chute de /r/

Kawaguchi (2009)

Corpus 2005-2006

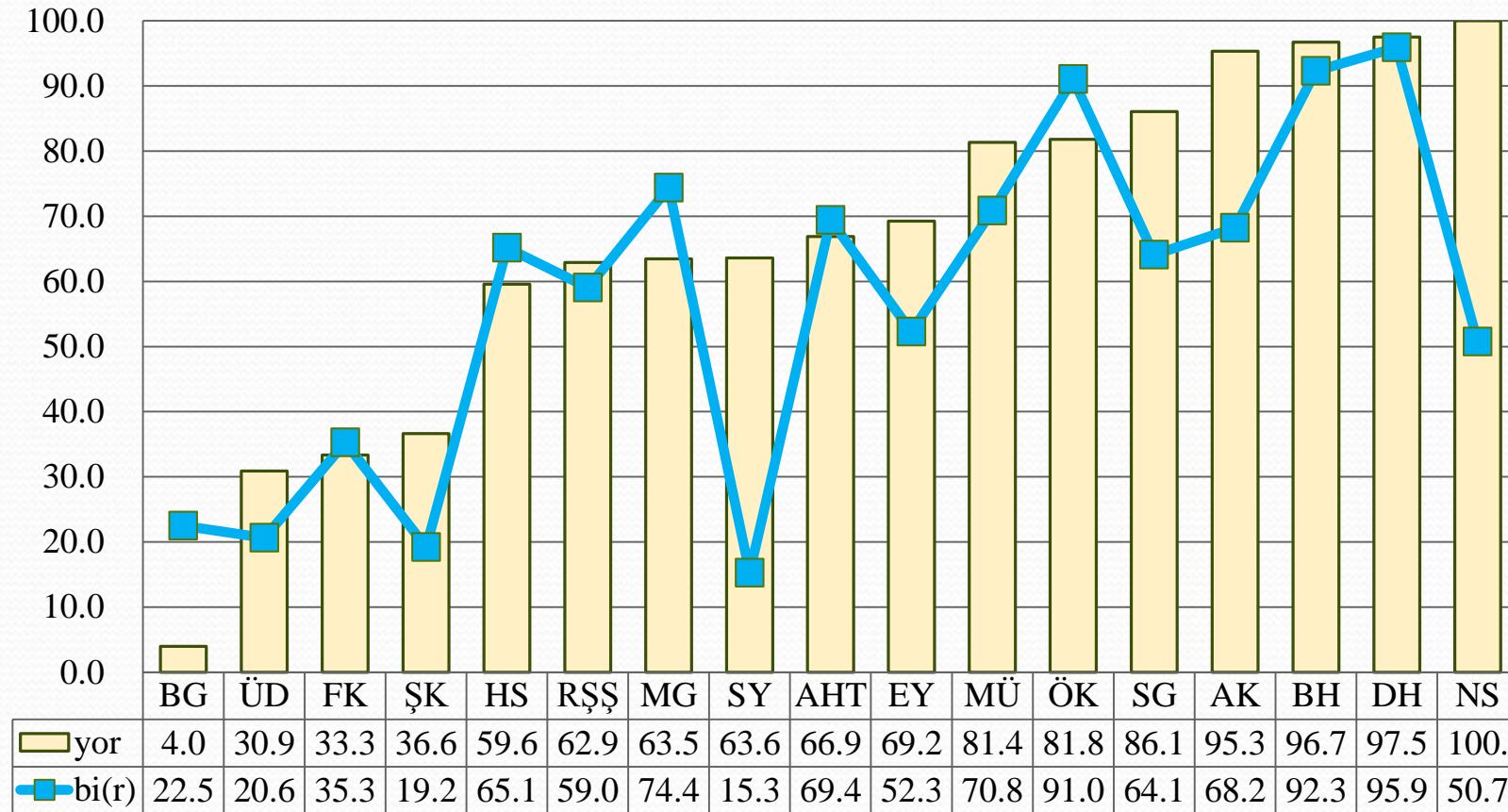
10h26m, mots types 17,417, mots 93,837

paramètres différents

- sujet de conversation : $r = 0.11$
- vitesse de conversation : $r = 0.25$
- différence sexuelle : $r = 0.16$
- différence d'âge: $r = - 0.19$

(r = coefficient de corrélation Pearson)

Variation individuelle de la chute de -r dans -Iyor et bi(r)



|-----|
moins de chute de -r

moyenne de la chute de -r :
-Iyor = 69.0 %; bir = 61.8 %

Conclusion

Analyses diverses basées sur le corpus de la langue parlée

- analyse structurale
 - variations phonologique, morphologique, syntaxiques
- analyses pragmatique et conversationnelle
 - analyse du discours, politesse, hésitation
- analyses sociolinguistique et stylistique
 - variations diastratique et diaphasique
- recherches appliquées
 - norme et usage, matériaux authentiques pour les apprenants

References

- Blanche-Benveniste, Claire. 1997. *Approches de la langue parlée en français*, Gap: Ophrys.
- 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: Routledge.
- Kawaguchi, Y., S. Yılmaz & A. Uras Yılmaz. 2006. “Intonation patterns of turkish interrogatives”, in *Prosody and syntax. Cross-linguistic perspectives. Usage-based linguistic informatics (UBLI) 3*, ed. Y. Kawaguchi, I. Fónagy and T. Moriguchi, Amsterdam/Philadelphia: John Benjamins Publishing Company, 349-368.
- 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.
- Özsoy, A. Sumru. 2004. *Türkçenin yapısı – I, Sesbilim*, İstanbul: Boğaziçi Üniversitesi Yayınları.
- Sarıca, M. (ed), 2005. *Sözlü Dil Yapısı : Yeni Dilbilim Kuramları Işığında*, ortak kitap, İstanbul: Multilingual.
- Uras Yılmaz, A. 2005. *Le fonctionnement des ligateurs dans l'échange discursif en français et en turc*, İstanbul Üniversitesi, Edebiyat Fakültesi, Dilbilim Dergisi, XIV, 87-107.
- Uras Yılmaz A., Yılmaz S. & Morel M.-A. (ed), 2004. *Vers une grammaire linguistique du turc*, ortak kitap, İstanbul: Multilingual.
- 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.

Merci de votre attention !

coelang@tufs.ac.jp

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers 24320102

Grant-in-Aid for Scientific Research(B) Responsible:Yuji KAWAGUCHI.