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

  Turkish National Corpus

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

- Corpus du turc parlé:

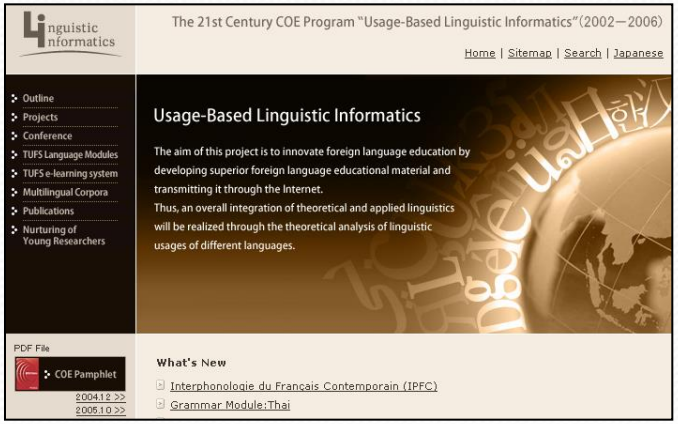
  Spoken Turkish Corpus (STC)

  http://std.metu.edu.tr/en

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

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

2.3. Corpus du turc parlé

21st Century COE Program

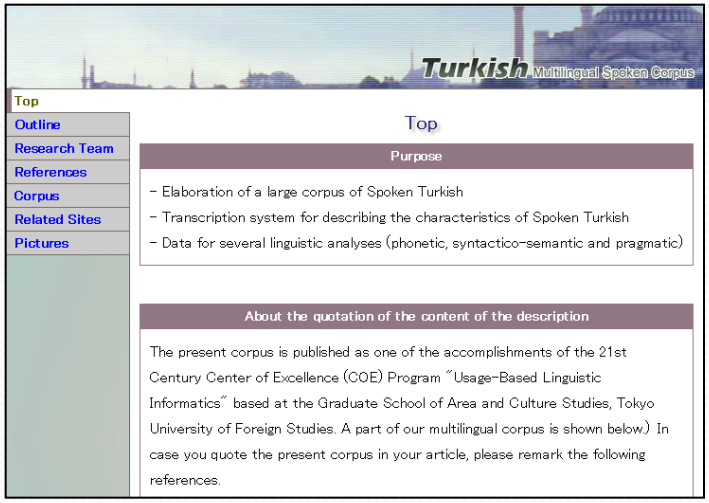
Selim Yılmaz (Université de Marmara),
Arsun-Uras Yılmaz (Université d'Istanbul)

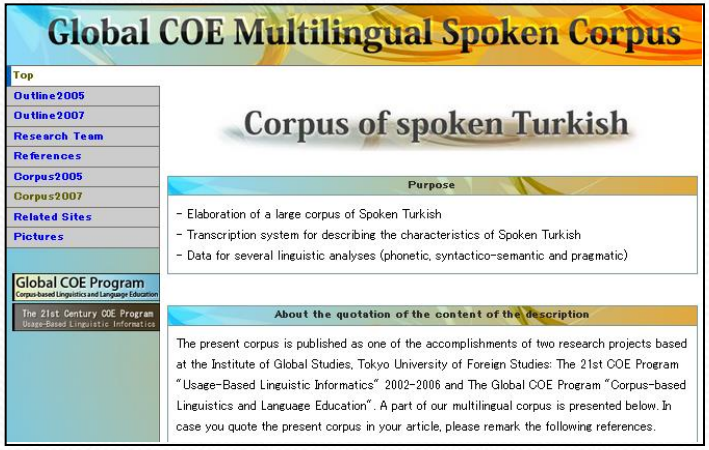
2005: 29 conversations : 98,100 mots

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





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
  → stuructures syntaxique/morphogique/ 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 “maladroit” ?
  → 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'informanteur

*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

Signe	Sens
Nom	Ex: SY = Selim Yılmaz. SY1, SY2, SY3
(?), (!)	Interrogation et exclamation
-	début du discours. Ex: -SY1....
#, ##, ###...	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:## ancak Kastamonu'nun çevresi ilçeleri çok güzel
e, ee, eee,…	hésitation Ex: yoksa ee yani çek(.) çekilmez 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: çünki 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 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 ilmi de o kadar yeterli değil # yani kendisi de bunun farkında #

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

.....

(2007-05)

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}

.....

5. Analyse basée sur le corpus
5.1. Structure de l'énoncé oral turc

MARQUES DISCURSIVES (Déictique / Embrayeur)	VALEURS / FONCTIONS ENONCIATIVES
YANİ	Explication / Explicitation / 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 a valeur contrastive / Antithese / Possibilité de discordance
BEN / BANA	Point de vue du sujet parlant (jugement personnel) / Assertion / Prise en charge avec position égocentrée

5.2. Chute de /r/

Kawaguchi (2009)

Corpus 2005-2006

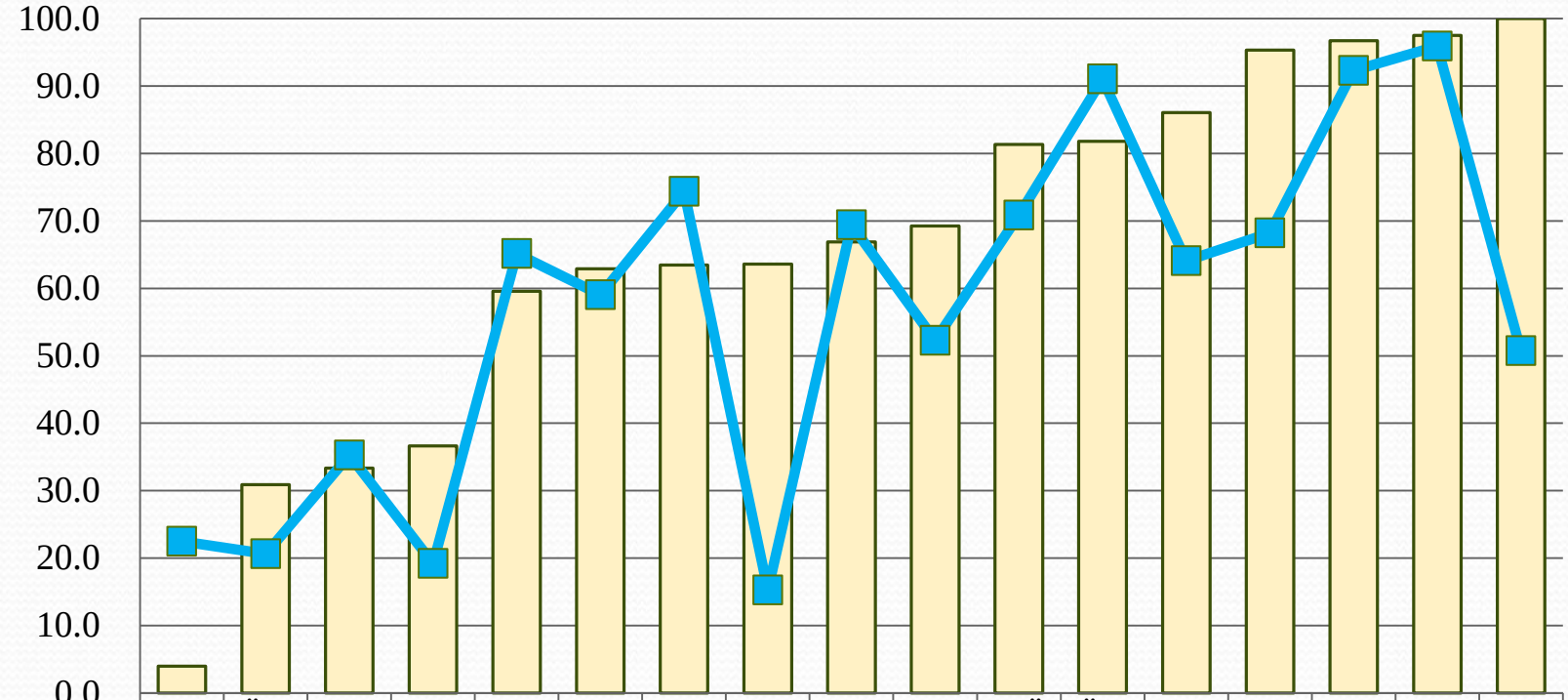
10h26m, mots types 17,417, mots 93,837

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 correlation Pearson)

Variation individuelle de la chute de -r dans *-Iyor* et *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

|----------|
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, hésiation
- analyses sociolinguistique et stylistique
  variations diastratique et diaphasique
- recherches appliequées
  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: Routlege.

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, Istanbul: 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.