

*Présentation du corpus oral en français de TUFS
et son application pour l'analyse linguistique*

Hisae AKIHIRO & Yuji KAWAGUCHI,
Tokyo University of Foreign Studies (TUFS)

Sommaire

I. Introduction

II. Cadre général du projet de TUFS

III. Corpus oral en français de TUFS

IV. Application pour l'analyse linguistique

I. Introducion

Situations en France

- Constitution et exploitation de corpus oraux
- Changement l'approche du langage
- Application à différentes disciplines

Situations au Japon

- Traitement automatique des langues avancé plutôt dans le domaine de sciences de technologie
- Mépris de l'importance de l'utilisation de corpus dans l'analyse linguistique
- Très peu de corpus constitués

II. Cadre général du projet de TUFS

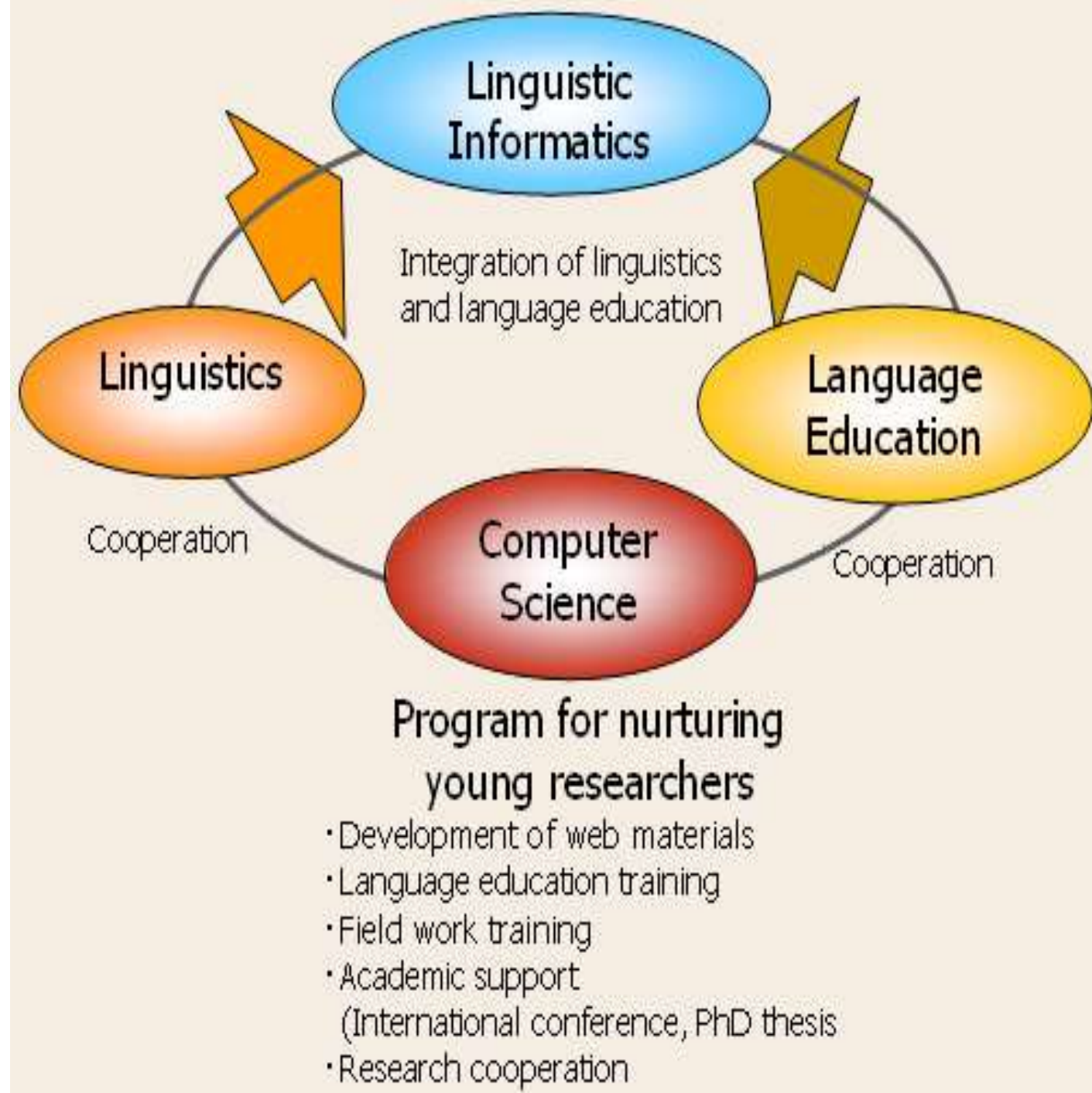
The 21st century COE Program 'Usage-Based Linguistic Informatics'
(2002-2006) <http://www.coelang.tufs.ac.jp/english/index.html>

Usage-Based Linguistic Informatics

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.

- Chef du projet : Y. KAWAGUCHI
- Corpus oraux de langues :
 - ❑ Français
 - ❑ Turquie
 - ❑ Malaisien
 - ❑ Espagnol
 - ❑ Canadien (de locuteurs bilingues)
 - ❑ Japonais (de locuteurs natifs)
 - ❑ Japonais (d'apprenants)



Global COE Program 'Corpus Based Linguistics and Language (2006-2012) Education' <http://cblle.tufts.ac.jp/index.php>

« The objective of this Global COE Program, 'Corpus-based Linguistics and Language Education' (CbLLE), is to foster internationally competitive researchers in subfields of linguistics that emphasize the empirical (particularly corpus-based) study of language structure and language education. Specifically, this COE project aims at training researchers and educators in acquiring a profound understanding of the nature and significance of linguistic and cultural diversity through the solid education program and active international academic exchanges at the Graduate School of Area and Culture Studies of TUFS. »

- Coordinateur du projet : Makoto Minegishi
- Corpus oraux de langues
 - ❑ Espagnol,
 - ❑ Français (de locuteurs natifs)
 - ❑ Français (d'apprenants)
 - ❑ Japonais (de locuteurs natifs)
 - ❑ Japonais (d'apprenants) en collaboration avec IPFC
 - ❑ Swahili
- Moteur de recherche 'POS' (Part of Speech)
 - ❑ Espagnol
 - ❑ Français
 - ❑ Japonais (d'apprenants)

III. Corpus français en oral de TUFS

- Superviseurs :
 - Yuji Kawaguchi (TUFS)
 - Hisae Akihiro (TUFS)
 - Atsushi Sano (Université de Fukushima)
- Organisateurs :
 - Kaori Sugiyama (Université de Seinan Gakuin)
 - Sunsuke Nakata (TUFS)
 - Mito Matsuzawa (TUFS)
 - Nori Kondo (TUFS)
 - Kentaro Koga (TUFS)
 - Misato Kikuchi (TUFS)
 - Françoise Lorant (Université de Paris XIII)
 - Takahiro Ogawa

- Collaborateurs :
 - †Claire Blanche-Benveniste (Université d'Aix-Marseille)
 - José Delofeu (Université d'Aix-Marseille)
 - Frédéric Sabio (Université d'Aix-Marseille)
 - André Valli (Université d'Aix-Marseille)
 - Jeanne-Marie Debaisieux (Université de Paris III)
 - Christophe Benzitoun (Université de Lorraine)
 - Takaaki Shochi (Université de Bordeaux III)
- Programmeurs :
 - Tsuyoshi Umeno (TUFS)
 - Kaori Omura (TUFS)
- Correcteurs de annotation morphosyntaxique :
 - Etudiants de Master de TUFS

Recueil des données

Année	Lieu	Nb d'enregistrement	Nb d'heures	Nb de mots	Type de données
2005	Aix-Marseille (Université d'Aix-Marseille)	21	7 h	158,763	Conversation libre Improvisation
2006	Paris (Université de Paris 13)	7	8h30	59,603	Conversation libre autour du sujet de la musique
2010	Aix-Marseille (Université d'Aix-Marseille)	34	30h	595,213	Conversation libre Interview
2011	Aix-Marseille (Université d'Aix-Marseille)	32	30h	En cours de transcription	Conversation libre
2013	Bordeaux/Paris (Université de Bordeaux, Paris 3, INALCO)	22	19h30	En cours de transcription	Conversation libre

Transcription

	2005	2010-
<i>Pause brève</i>	+	#
<i>Pause longue</i>	++	##, ###
<i>Voyelle longue</i>	quoi: eh:	Quoooooi, Eeeh
<i>Amorce d'un mot</i>	mer- (<i>mercredi</i>)	mer(.) (<i>mercredi</i>)
<i>Chevauchement de paroles</i>	_____	< >
<i>Bruits non verbaux</i>	{rire}	[rire]
<i>euh</i>		e, ee, eee
<i>? / !</i>		?/!
<i>hum</i>		mm, mm, mmm

	2005	2010-
<i>Élément disloqué</i>		Il vous va très bien, {ce pantalon}
<i>Mise en accent</i>		_____
<i>Ajout d'un schwa</i>	donc° sur°	
<i>Interruption</i>	///	
<i>Multi-transcription</i>	/j'ai été, j'étais/	
<i>Hésitation orthographique</i>	il(s) travaille(nt)	
<i>Manque de liaison</i>	vous # allez	
<i>Incompréhensible</i>	X, XX, XXX	X, XX, XXX
<i>Locuteurs</i>	L1, L2, ... fr01, fr02,...	Jean Dubois = JD1, JD2,...

Exemple de transcription 2005

Fr01 – bonjour

Fr02 – bonjour {sourire} ++

Fr01 – je pars en voyage d'affaire euh à Mexico je dois être là-bas
lun(di)

Fr02 – bon je vais regarder + mais je vous XX

Fr01 – je ne peux pas je ne peux pas partir avant samedi

Fr02 – mais je vous préviens que ça va être assez difficile

Fr01 – hum / d'accord, ø / ++

Fr02 – il n'y a que trois vols Air France par semaine

Fr01 – que- {raclement de gorge} quels jours

Fr02 – le lundi le mar- pardon euh + le lundi le mercredi et le vendredi

Fr01 – la compagnie ça m'est complètement égal

Fr02 – Air Mexique ou Trans – XXXXX {bruit}

Exemple de transcription 2010

AM1 – Bonjour.

DH1 – Bonjour. [sourire] ###

AM2 – Je pars en voyage d'affaire ee à Mexico. Je dois être là-bas lun(di).

DH2 – Bon je vais regarder #

AM3 – Je ne peux pas <mais je vois XX>, je ne peux pas partir avant samedi.

DH3 – Mais je vous préviens que ça va être assez difficile.

AM4 – mm. xx ##

DH4 – Il n y a que trois vols Air France par semaine.

AM5 – Que(.) [raclement de gorge] quels jours?

DH5 – Le lundi, le mar(.) pardon, euh # le lundi, le mercredi et le vendredi.

AM6 – {La compagnie} ça m'est complètement égal.

AM7 – Air Mexique ou Trans-XXXXX [bruit]

Démonstration de corpus

http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.html?contents_xml=corpus&menulang=en



The screenshot displays the website for the French (Aix) Multilingual Spoken Corpus. At the top, there is a banner image of a university campus with a river, overlaid with the text "French (Aix) Multilingual Spoken Corpus". Below the banner, a navigation menu on the left lists several options: "Top", "Outline", "Research Team", "References", "Corpus" (which is highlighted), "Related Sites", and "Pictures". The main content area, titled "Corpus", contains a list of four items, each with a red circular icon to its right:

1. Student affairs section
2. Reception of the university
3. About faculty
4. About health condition

Usage-Based Linguistic Informatics | Language Function-Specific Corpus

JA | HOME > Corpus > Student affairs section

French (Aix) Multilingual Spoken Corpus

[Top](#)
[Outline](#)
[Research Team](#)
[References](#)
Corpus
[Related Sites](#)
[Pictures](#)



☒ French
☐ Japanese
☐ English

☐ Scroll On/Off

Student affairs section

French; L1 voilà donc ton rôle

French; L2 euh vice-président étudiant donc +

French; euh donc dans l'université il y a quatre vice-présidents +

French; trois vice-présidents qui sont des professeurs

French; et un autre qui est un vice-président étudiant

French; L1 d'accord

French; L2 pas tout à fait au même titre /parce que, XX/ pour l'instant les statuts de l'université ne sont pas encore euh clarifiés par rapport à ça

French; c'est quelque chose d'assez nouveau

French; L1 mh mh

French; L2 dans l'université et dans les université françaises en général(e)

French; donc pour l'instant j'ai un statut de chargé de mission du président

French; mais à la rentrée prochaine normalement les statuts devraient changer pour que

vice-président étudiant devienne un vice-président à part entière en gros < pour

IV. Application pour l'analyse linguistique

Analyse morphosyntaxique

Moteur de recherche 'POS (Part of Speech)'

<http://cblle.tufts.ac.jp/tag/fr/index.php?menulang=en>

'Tree tag' conçu par Schmid, H (1995)

<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

POS Search Engine ➡

■ What's New

• (218366 words)

■ Outline & Purpose

CbLLE POS Research Engine (Spoken French) is to search part-of-speech tags of the *Multilingual Spoken Language Corpora (French)* developed by the 21st Century COE Program “Usage-Based Linguistic Informatics”. Part-of-speech (POS) tagging is based on the system called *Tree Tagger*, cf. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>. The purpose of this research engine is largely to publish the research results from the Global COE Program “Corpus-based Linguistics and Language Education (CbLLE)” and to analyze various grammatical phenomena in spoken French.

■ Corpus Description

The current corpora of Spoken French are composed of two different datas.

1. Twenty dialogs recorded during July 2005 with the cooperation of the University of Aix-Marseille I. The total length is around seven hours and the total number of word tokens exceeds 100,000. They cover various subjects such as office life, work at the university, recent news about one's family and improvised sketches.

References

http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.php?contents_xml=gaisetsu&lang=ja

http://www.coelang.tufs.ac.jp/multilingual_corpus/fr2/index.php?contents_xml=gaisetsu&lang=ja

pos Any:POS KON Any:POS Search 5 5 Français JA HELP search sample
lemma savoir Reset
token
☐ Ignore the second context

Conserve results Conserve results with contexts

No of sentences :68
No of tokens :68

- 03_MW_CD_100202.txt:25090 Sachant que en début juillet il y a autre chose que la Japan Expo n'est-ce pas ? [rires]
- fr03-1.txt:9506 de de de savoir s' ils ont un projet eux-mêmes
- fr03-1.txt:9656 moi je sais que je suis passé par là > et
- fr03-1.txt:9688 oui > sachant qu' ils doivent pas les payer en fait
- fr03-3.txt:10425 euh tout ça les étudiants le savent et +
- fr03-3.txt:10455 ouais mais bon si enfin tu savais que à sept heures il y a un truc super intéressant à la fac
- fr03-3.txt:10522 faut savoir qu' il y a trente pour cent euh
- fr03-3.txt:10571 donc maintenant je sais que je vais faire ça et ça
- fr04.txt:10904 bon mon ancien directeur le sait que je déborde toujours et que je m'occupe beaucoup des < étudiants
- fr04.txt:10934 et bon alors tu sais que > c' était < c' était important hein
- fr04.txt:10995 et ce travail me convient bien parce qu' il sait que je suis autonome
- fr04.txt:11476 je sais que les étudiants le rendent bien
- fr04.txt:11548 ça Olivier > lui c' est pour euh les étudiants tu sais quand ils ont leurs notes et qu' il y a un problème qui va pas < XX
- fr05-1.txt:11827 je veux savoir si tu fatiguée
- fr05-2.txt:12223 ouais mais bon tu sais que c' est toi quoi < c' est
- fr05-2.txt:12310 pour savoir si tu es pas cardiaque
- fr05-3.txt:12560 tu sais qu' il y a des gens qui vont l' entendre demain +
- fr09-1.txt:12582 je sais que c' est voyage plutôt organisé + en gros
- fr12.txt:13335 et euh donc très bien tu fais ça mais ensuite / il , 0 / faut savoir que la plupart des étudiants de classe prépa ensuite se dirige vers l' université +
- fr12.txt:13498 ouais le contenu des cours euh ils te disent grosso modo euh ce qu' on va faire quoi par exemple en linguistique je sais qu' on aura à faire +
- fr13.txt:13605 je sais que ma soeur elle avait terminé sa maîtrise elle avait mis un an et demi pour faire sa maîtrise +

COLLE POS Research E...

cbille.tufs.ac.jp/tag/fr/search.php?menulang=en

Google

よく見るページ Firefox を使いこなそう Web スライス ギャ... おすすめサイト

pos Any.POS KON Any.POS

lemma savoir

token

Search Reset

Ignore the second context

Français

ja HELP search sample

5 5

paris

aix 2005

aix 2010

Conserve results

Conserve results with contexts

No of sentences :68

No of tokens :68

1. Oui . [rires]

2. Comme ça on pourrait voir tous les concerts, ça serait vraiment bien.

3. Oui.

4. Mais, en fait ils n'ont pas fait # le planning.

5. Il me tarde de savoir qui c'est qui va venir !

03_MW_CD_100202.txt:25090 Sachant que en début juillet il y a autre chose que la Japan Expo n'est -ce pas ? [rires]

1. Yeahhh !

2. Donc e.

3. Non c'est fin juillet ! Pourquoi tu oublies à chaque fois ?

4. Non, non, c'est le 6.

5. le 6 ?

fr03-1.txt:9506 de de de savoir s' ils ont un projet eux-mêmes

fr03-1.txt:9656 moi je sais que je suis passé par là > et

fr03-1.txt:9888 oui > sachant qu' ils doivent pas les payer en fait

fr03-3.txt:10425 euh tout ça les étudiants le savent et +

fr03-3.txt:10455 ouais mais bon si enfin tu savais que à sept heures il y a un truc super intéressant à la fac

fr03-3.txt:10522 faut savoir qu' il y a trente pour cent euh

fr03-3.txt:10571 donc maintenant je sais que je vais faire ça et ça

fr04.txt:10904 bon mon ancien directeur le sait que je déborde toujours et que je m' occupe beaucoup des < étudiants

fr04.txt:10934 et bon alors tu sais que > c' était < c' était important hein

fr04.txt:10995 et ce travail me convient bien parce qu' il sait que je suis autonome

fr04.txt:11476 je sais que les étudiants le rendent bien

fr04.txt:11548 ça Olivier > lui c' est pour euh les étudiants tu sais quand ils ont leurs notes et qu' il y a un problème qui va pas < XX

fr05-1.txt:11827 je veux savoir si tu fatiguée

fr05-2.txt:12223 ouais mais bon tu sais que c' est toi quoi < c' est

fr05-2.txt:12310 pour savoir si tu es pas cardiaque

fr05-3.txt:12560 tu sais qu' il y a des gens qui vont l' entendre demain +

fr09-1.txt:12582 je sais que c' est voyage plutôt organisé + en gros

fr12.txt:13335 et euh donc très bien tu fais ça mais ensuite / il , 0 / faut savoir que la plupart des étudiants de classe prépa ensuite se dirige vers l' université +

fr12.txt:13498 ouais le contenu des cours euh ils te disent grosso modo euh ce qu' on va faire

pos Any:POS DET:ART NOM Search

lemma savoir Reset

token

☐ Ignore the second context

Français

5 5

paris

aix 2005

aix 2010

search sample

Conserve results Conserve results with contexts

No of sentences :10
No of tokens :10

1. fr03-3.txt:10494 de de **savoir les implications** enfin
2. fr05-2.txt:12267 tu **sais le film** là < où en fait tu t' amuses à faire
3. fr13.txt:14001 tu **sais des fois** l' article défini l' article indéfini ça change beaucoup de < sens
4. fr14.txt:14648 tu **sais un chèque** que j' avais dû prendre comme ça pour le cas où il aurait fallu que je retire de l' argent +
5. fr15.txt:16308 parce que bon euh tu **sais le pétrole** et tout ça bon peut-être que il y aura des X +
6. fr16.txt:17149 tu as **su les résultats**
7. fr16.txt:17513 XXX tu **sais l' escalier** qui montait euh
8. fr29.txt:19666 tu **sais les combinaisons** là +
9. fr52.txt:22540 oui ben à un moment donné tu **sais les gens** ils s' usent et là c' est un effet de masse + tu as un référendum qui est pas passé une première fois après ça peut + au contraire enfoncer le clou + mais euh ils sont forts tous les médias qui sont pour conditionner les gens euh +
10. fr57.txt:1886 vous **savez les loyers** en ville sont chers

pos Any:POS VER:infi Any:POS Search

lemma savoir Reset

token Ignore the second context

5 5

paris

☒ aix 2005

☒ aix 2010



Conserve results Conserve results with contexts

No of sentences :1
No of tokens :1

1. fr21.txt:18376 * mais c' est lui qui sait faire aussi hein

Etudes basées sur le corpus français de TUFS

- Interface entre la phonologie et la syntaxe

Nakata & Kawaguchi (2005)

Kondo (2010/2012)

- Syntaxe (micro et macro)

Akihiro (2006)

Iida (2013)

- Enseignement du français pour les étudiants japonais

Sugiyama (2011, 2012)

Bibliographies

Cappeau, P. & Seijido, M. (2005). Les corpus oraux en français (inventaire 2005 v.)
http://www.dglf.culture.gouv.fr/recherche/corpus_parole/Presentation_Inventaire.pdf

Akihiro, H. (2006). 「フランス語会話コーパスにおける直接目的語省略の問題について」(à propos de l'ellipse de l'objet direct dans un corpus de conversation dialogique en français), 『言語情報学研究報告12』(Rapports de Linguistique informatique 12), Tokyo, TUFs:183-199.

Iida, R. (2013). フランス語の話し言葉における従属接続詞 *parce que* の機能的拡張—日本語の「カラ」との比較を通して (Elargissement de la fonction de la conjonctive de subordination '*parce que*', dans le français parlé – en comparaison de '*kara*' en japonais) Mémoire de Master, Nagoya, Université de Nagoya.

Koga, K. Akihiro, H. Kawaguchi, Y. (2011) 「Aix話し言葉コーパスプロジェクト」(Projet du corpus oral d'Aix) Flambeau 37, TUFs : 37-54.

Kondo, N. (2010). 「フランス語の話し言葉におけるリエゾン—Aix-en-Provence コーパスを用いた統語論・音韻分析—」(Liaison dans le français parlé – analyse syntaxique et phonologique en utilisant de corpus d'Aix-en-Provence), Mémoire de Master, Tokyo, TUFs.

Kondo, N. (2011). 「フランス語のリエゾンにおける社会言語学的要因の一考察」(Réflexion sur des facteurs sociolinguistiques de variation de la liaison dans le français parlé). 「ロマンス語研究44」(Etudes Romaines 44), Tokyo, Société japonaise des études romaines: 59-68

Nakata, S. & Kawaguchi, Y. (2005). 「統語構造を反映する韻律特徴—文の意味とイントネーション, 音節長, ポーズ—」Traits prosodiques à la lumière de structures syntaxiques – Significations de phrases, intonation, durée syllabique et pause. 『フランス語を探る フランス語学の諸問題III』*Explorer le français. Problèmes de la linguistique française III*, Tokyo, Sanshûsha: 308-321.

Sugiyama, K. (2011). 「フランス語初級学習者の話し言葉における使用語彙の複雑さ」(Caractère complexe du lexical utilisé dans la production orale chez les apprenants débutants de français) 『ロマンス語研究44』 (Etudes Romaines 44), Tokyo, Société japonaise des études romaines: 49-58.

Sugiyama, K. (2012). Lexical Profile of French Learner Speech: The Case of Japanese University Students. In Tono, Y., Kawaguchi, Y. & Minegishi, M. (éds.), *Developmental and Cross linguistic Perspectives in Learner Corpus Research* Amsterdam/Philadelphia, John Benjamins.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 26370476, 24320102.

Grant-in-Aid for Scientific Research(C)

Person in charge:Hisae AKIHIRO.

Grant-in-Aid for Scientific Research(B)

Person in charge:Yuji KAWAGUCHI.