学習者と母語話者の使用語彙の違い
――『日中Skype会話コーパス』を用いて

中俣尚己

要旨

『日中Skype会話コーパス』は真正性のある接触場面会話コーパスである。このコーパスを用いて、学習者と母語話者の使用語彙を比較したところ、大局的な視点では学習者と母語話者のTTRに違いはなく、語彙の豊富さは同じであるという結論が得られた。次に、品詞別に分析をしたところ、実質語ではほとんど差がなく、機能語、特に副詞に差が見られるという結果が得られた。副詞は学習者の使用は母語話者の6割にとどまる反面、学習者は多用するが、母語話者はほとんど使用しないという副詞もあり、ズレが大きいことが確認された。このようなコーパスから得られた知見を、教育に反映していく必要がある。

キーワード
接触場面、意味交渉、TTR、特徴語、対数尤度比（LLR）、副詞

Abstract

This article clarifies the differences between the vocabulary used by Non-Native Speakers (NNS) and Native Speakers (NS) that appeared in a "Japan-China Skype-Conversation Corpus" that the author created. The corpus contains authentic Japanese conversations in contact situations.

From a macro-perspective, there are no differences between the TTR of NNS and NS, which means that there is no divergence found in vocabulary. Next, considering the parts of speech, no significant differences are found in substantial words such as nouns and verbs. However, function words, especially adverbs, show great differences. The token frequency of adverbs used by NNS is as small as 60% of those used by NS. However, some adverbs are mainly used by NNS, while being rarely used by NS.

Key words
Contact Situation, Meaning Negotiation, TTR, Specific Words, Log Likelihood Ratio (LLR), Adverb

Differences Between the Vocabulary Used by Non-Native Speakers and Native Speakers
Using a Japan-China Skype-Conversation Corpus

Naoki Nakamata

1 はじめに

この論文では、まず筆者が構築した接触場面会話コーパスである『日中Skype会話コーパス』の特性について述べる。このコーパスを使った研究は様々な形で行うことができるが、今回は真正性を持った接触場面であるという点を重視し、学習者と母語話者の使用語彙についてマクロな観点から比較する。特に品詞によって差異が見られるのかを明らかにし、なぜ差異が見られたり見られなかったりするのかを考察する。

2 『日中Skype会話コーパス』について

2.1『日中Skype会話コーパス』の概要

『日中Skype会話コーパス』は2012年5月～7月に、東京・実践女子大学と長沙・湖南大学の学生間で行ったSkypeを利用した遠隔会話活動（中俣ほか2013）を録音、文字化したもので、接触場面の日本語会話コーパスに分類される。中国側の学習者は全員2年生である。日本側の母語話者は学部3年～大学院博士前期課程1年の学生で日本語教育を専攻したり、関連する授業を受講していた学生である。3ヶ月の間、ペアを固定し、1週間に1度のペースでSkypeを用いた日本語会話活動を行った。1回の会話は最大90分を目安とした。実際にはビデオ通話ではあるが、行ったのは録音のみで、現時点で公開しているのはその文字化資料のみとなる。コーパスはテキスト形式で、筆者のホームページから、無料でダウンロード可能である[注1]。

コーパスには延べ9ペア、38の会話を収録している。総会話時間は46:48:35で、1会話あたり平均1:13:55とまとまった長さの会話と言える。後述する日本語解析システム「雪だるま」を使って分析した結果、総語数は204,632語であった（記号類を除く）。

2.2『日中Skype会話コーパス』の特性

『日中Skype会話コーパス』の言語資料としての特徴として、「真正性」「縦断的」「電話場面」「話題の指定」の4つを挙げることができる。

2.2.1 真正性がある

このコーパスの設計はもともとコーパスを作ろうとしたものではなく、まずはSkypeを用いた会話活動を通し、中国の学習者には学んだ日本語を使う機会を提供するとともに学習意欲を継続させること、日本の母語話者には外国人と文化交流をしたり日本語を教えたりしながら、日本語について考えてもらうことが第一の目的であり、それにあわせてSkypeを用いた会話活動の計画がデザインされている。そのため、真正性のある接触場面コーパスになっている。以下、いくつかの語について、代表的な学習者コーパスであるKYコーパスと比較したものが表1である。OPIという統制された会話であるKYコーパスには出現しないような語が『日中Skype会話コーパス』には多数出現していることがわかる。

表1 KYコーパスと日中Skype会話コーパスの出現数の比較[注2]

語	KYコーパス	日中Skype会話コーパス
明後日	0	7
木曜	6	41
すごい	77	211
すごく	190	86
すげえ	0	4

2.2.2 縦断コーパスである

会話活動は1週間に1回、継続的に行った。最も多いペアで7回分の会話があり、縦断的にデータを観察することができる。

2.2.3 一種の電話場面である

終結部には、例えば突然食事の話題をふって、会話を終結にもっていく前終結の段階が存在するなど、電話場面と同様の構造が観察される（橋内1999）。ま

た、コミュニケーション・ブレイクダウンや沈黙も多く観察される。

2.2.4 話題が指定されている

各回は表2のように話題が指定されており、数字はファイル名の末尾の数字に対応する。しかし、話題は必ずしも厳密に守られているわけではなく、話がそれたり日本語についての質問が行われることもある。これらの話題は事前に日中双方の学生から話してみたいことのアンケートを行い、決定した。

表2 日中Skype会話コーパスの話題

1	ポップカルチャー	6	伝統・行事
2	料理	7	夏休み・夏の予定
3	家庭・家族・子供	8	大学生活
4	故郷・今住んでいる場所	0	指定なし・トピック認定できず
5	敬語		

3 学習者と母語話者の使用語彙の違い

3.1 語彙量の違い

様々な観点から分析が可能な『日中Skype会話コーパス』であるが、本研究では端緒としてマクロ的な視点から学習者と母語話者の使用語彙の違いを明らかにする。まず、コーパスを学習者の発話と母語話者の発話に分割し、それぞれを日本語単語解析システム「雪だるま」で単語に分割した[注3]。

この「雪だるま」は長岡技術科学大学の山本和英氏が開発したシステムで、形態素ではなく「単語」に分割することを目的とし、「気が早い」のような慣用句、「かもしれない」のような複合辞、「勉強する」のようなサ変動詞、「無理だ」のような形容動詞をそれぞれ1語として出力することができる。解析は2015年9月7日に行った。

まず、学習者と母語話者の全発話のToken頻度、Type頻度ならびにそこから産出されるTTRを表3に示す。TTRとはType頻度をToken頻度で割った値で、語彙の豊富さを計量するのに広く使われている指標である（石川2012）。

表3 学習者と母語話者の使用語彙量の比較

	学習者	母語話者
Token頻度	104,156	100,325
Type頻度	5,434	5,217
TTR	0.052	0.052

表3から明らかなように、学習者と母語話者のToken頻度、Type頻度はほぼ等しく、TTRにも全く差は見られない。発話量、語彙の豊富さともに学習者と母語話者の間に違いはないということである[注4]。これは本コーパスの特徴というよりは、真正性のある接触場面の特徴と言えるかもしれない。別の真正性のある接触場面の発話を形態素解析して調査した宇佐美・中俣（2013）も学習者と母語話者の使用語彙に差は見られなかったと報告している。

3.2 品詞ごとの違い

3.1では発話されたすべての語をまとめて計量を行ったが、品詞ごとに見るとどのような違いがあるだろうか。図1は、各品詞ごとに、母語話者のToken頻度を1とした時の学習者のToken頻度を示したものである[注5]。また、表4に品詞ごとのToken頻度の実数を示す。

図1からわかるように、学習者と母語話者の違いが大きいのは副詞、助動詞、接続詞といった機能語の類である[注6]。動詞も差が大きいが、これは仔細に分析すると「田中という人」のような「言う」の使用に大きな開きが見られることが最大の要因である。この「言う」はむしろ機能語的に使われているものと言えよう。対して、名詞や形容詞といった実質語では違いは大きくない。これは全数の割合だけでなく、頻度リストを作ってみても、母語話者は多用するのに学習者があまり使わないといった語が見つからない、という結果になっている。

なお、学習者が感動詞を多用しているが、これは外国語を話すのに伴ってフィラーの発話が増えたためである。このことは至極当たり前の現象であり、注目には値しないだろう。

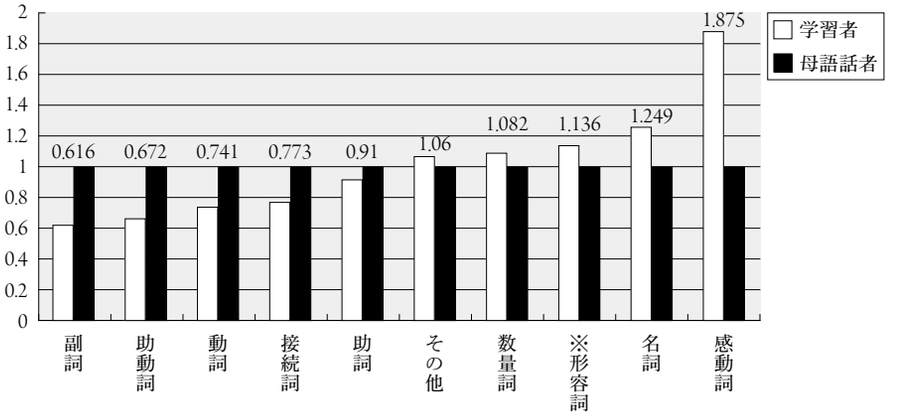

図1　品詞ごとの母語話者を1とした時の学習者のToken頻度の割合

表4　品詞ごとの学習者と母語話者のToken頻度

品詞	学習者	母語話者	品詞	学習者	母語話者
副詞	3,477	5,646	その他	1,389	1,310
助動詞	9,693	14,421	数量詞	1,416	1,309
動詞	7,807	10,531	形容詞	7,002	6,163
接続詞	578	748	名詞	23,853	19,098
助詞	26,510	29,133	感動詞	22,431	11,966

4 副詞に見られる学習者と母語話者のズレ

3節で、機能語において学習者と母語話者に違いが見られることを確認した。ここでは、紙幅の都合から、最も大きなズレが見られた副詞に焦点を当てて論じることにする。

副詞の興味深い点は、Token頻度で見ると学習者は母語話者の6割程度しか使っておらず、学習者の発話には副詞が少ないのであるが、個々の語に注目すると、学習者が多用し、母語話者はほとんど使用しないようなものも多く見られるという点である。単なる数の問題ではない、大きなズレが見られるということである。

まず、副詞で使用頻度が多かった語を表5にあげる。NNSは学習者、NSは母語話者を意味する。

表5を見ればわかる通り、学習者（NNS）と母語話者（NS）に差がある語が多い。だいたいは母語話者の方が多いが、「とても」のように学習者が母語話者の8倍も使用しているような語もある。「一番」「もし」「あまり」なども学習者の方がよく使用している。一方で、「結構」は学習者の使用数は少なく、使用数は母語話者の15分の1に満たない。

表5　『日中Skype会話コーパス』に多い副詞と使用者

副詞	NNS	NS	合計	副詞	NNS	NS	合計
そう	623	1,842	2,465	とても	159	19	178
ちょっと	346	595	941	一番	104	53	157
もう	142	327	469	結構	8	133	141
こう	34	303	337	いろいろ	95	37	132
どう	122	215	337	まだ	44	74	118
多分	172	118	290	もし	77	27	104
よく	84	140	224	やっぱり	21	77	98
あんまり	102	119	221	あまり	59	32	91
まあ	60	127	187	いっぱい	21	60	81
例えば	96	91	187	もっと	48	32	80

各話者に特徴的な語をあぶりだすために、すべての副詞について対数尤度比（Log Likelihood Ratio）を計算した。対数尤度比は特徴語の指標としてよく用いられており、教科書コーパスを利用した教科特徴語の抽出などもこれを使って行われている（田中・近藤2011）。計算方法は以下の通りである。

$$2(a\ln a+b\ln b+c\ln c+d\ln d-(a+b)\ln(a+b)-(a+c)\ln(a+c)-(b+d)\ln(b+d)$$
$$-(c+d)\ln(c+d)+(a+b+c+d)\ln(a+b+c+d))$$

a：当該資料での当該語の度数　b：参照資料での当該語の度数
c：当該資料の延べ語数－a　d：参照資料の延べ語数－b

lnは自然対数を表す。aまたはbが0の場合、alnaまたはblnbを0として計算する。
ad–bc<0の場合の場合、–1を乗じる補正を行う。

学習者と母語話者それぞれに特徴的な副詞を表6と表7に示す。絶対値が10以上の語は学習者で16語、母語話者に8語あった。

表6 学習者に特徴的な副詞

副詞	NNS	NS	LLR	副詞	NNS	NS	LLR
とても	159	19	208.0	あまり	59	32	26.8
いろいろ	95	37	63.1	きっと	26	6	25.1
つまり	35	1	59.6	特に	19	5	16.9
もし	77	27	56.0	もっと	48	32	15.8
たぶん	172	118	55.0	例えば	96	91	13.7
一番	104	53	51.7	ますます	6	0	11.6
ずっと	54	16	44.6	わざわざ	6	0	11.6
ただ	142	7	27.3	だいぶ	8	1	10.1

表7 母語話者に特徴的な副詞

副詞	NNS	NS	LLR	副詞	NNS	NS	LLR
そう	623	1,842	–245.5	なるほど	9	63	–23.7
こう	34	303	–140.3	ちゃんと	2	36	–22.8
結構	8	133	–82.6	取りあえず	0	16	–15.4
やっぱり	21	101	–25.7	もう	142	327	–13.3

このような差異が見られるのは単一の要因によるものではないだろう。例えば、母語話者に「そう」が多いのは「そうですね」「そうなんだ」のようにあいづちをうつ回数が母語話者が多かったことに起因する。一方で、学習者に「たぶん」「もし」のような呼応の副詞が多いのは、その文が非現実であることを早めに伝えてしまおうという気持ちの表れとも考えられる。

以下、強調表現として「とても」と「結構」を例に取り上げる。

（1）C：あ、長沙、長沙を紹介しよ。えーと、長沙は、〈はい〉えーと、とても広いですよ｛笑｝。 (02-06-4)

（2）J：うん。何時ぐらいに寝るんですか。
C：えっと、11時半ぐらいです。
J：ああ、そうなんだ。日本の大学生は結構みんな遅いと思います。 (01-05-1)

（1）は内容伝達の面では問題ないが、いかにも非母語話者の発話という印象を受ける。これを「結構広いですよ」のように言えるようになるとかなり「日本語ができる」という印象を与えられるのではないだろうか。ただし、母語話者が最も多用した強調表現は実は副詞ではなく、以下のようなものである。

（3）J：うん、あの、人気なんですけど、あの、アニメのその、映画？
C：アニメ、アニメの［映画。
J：［映画を作ってる人で、えっと、1年に1本ぐらい公開されていて、あの、すごく人気ですよ。 (01-05-1)

（4）J：うーん、〈ああ〉ピカチュウかわいいですよね｛笑｝。〈うんうん〉すごいかわいいキャラクターがいっぱい出てたので、〈うん〉人気でした。［●、うーんと、『ポケットモンスター』。 (03-05-1)

母語話者は「すごく」を76回、「すごい」を204回使用している。学習者の使用はそれぞれ13回、12回にすぎない。学習者の「すごい」は形容詞述語としての用法のみで、誤用とされる連用修飾用法は見られなかったが、母語話者は（4）のような連用修飾用法が半分程度を占めている。「とても」と「すごく」、そして「すごい」の連用修飾用法を加算した数は学習者と母語話者でほぼ等しくなることから、強調の副詞として何を使うかという選択の問題であることがわかる。

学習者が「とても」だけを使ってしまう要因として、「最初に教えられたものをつい使ってしまう」ということが考えられる。そうであるならば、初級教材においてどのような副詞を使うかということは非常に重要な問題である。

「とても」は無難な副詞ではあるかもしれないが、本コーパスのデータを見る限り、母語話者の「とても」使用は非常に少ない。また、その大半がエコー発話であった。さらに、フォーリナートークの影響も考慮に入れる必要があるだろう。初級から強調表現として「すごく」を導入してもそれほどデメリットはないのではないだろうか。

5 なぜ実質語に差が見られないのか

副詞とは対照的に、実質語には学習者と母語話者で大きな使用語彙のズレは見られなかった。母語話者が使う語は、学習者も使うのである。以下、よく使われる名詞20を挙げる。

表8 『日中Skype会話コーパス』に多い名詞と使用者

名詞	NNS	NS	合計	名詞	NNS	NS	合計
私	934	471	1,405	敬語	159	196	355
日本	750	465	1,215	自分	220	122	342
中国	603	256	859	料理	164	116	280
何	224	529	753	学校	122	123	245
人	332	408	740	いつ	163	72	235
こと	327	320	647	感じ	81	123	204
それ	283	302	585	授業	112	76	188
これ	192	306	498	大学	78	96	174
なん	165	198	363	一緒	108	66	174
先生	203	155	358	家	101	71	172

表8を見ると学習者と母語話者で大きく差がついている語がほとんどないことがわかる。これより下位の語もこの傾向は変わらない。「私」のみ学習者が2倍使っているが、これは一人称を表示するか否かという問題で、名詞の習得とは異なる問題であろう。

では、名詞ではなぜ学習者と母語話者に差が見られないのか。その原因の一つとして、名詞や動詞の場合、学習者と母語話者の知識にギャップがあっても、接触場面でそれが補正される、換言すれば意味交渉（Long 1983）を通じた学習が起こるということが考えられる。例えば、以下は「パスタ」という語を学習者が使った場面である。

（5）J：（略）パスタ、日本料理じゃなくて、パスタ。〈うーん〉パスタわかります？　パスタ。
C：さった？　［サラダ？
J：［パスタ。〈うん〉パスタ。
C：んー？
J：待ってね。《ポーズ　8秒》日本料理じゃないですよ、もう。パスタ。
C：ああ、パスタ。〈うん〉ピザですね。ん？　ピザです。
J：あの、麺。ピザとはちょっと違うかな。〈うーん〉麺、〈麺〉のパスタとか。［うん。
C：ああー。
J：ピザとか、お昼とかはそういうのみんな食べますね。今の若い子たちは。
C：このパスタは〈うん〉食材はなんですか。うーん。
J：麺です。パスタ。パスタ。〈麺〉パスタ。知らない？　うん。［あの。
C：［麺ですか。ああー。　（03-04-2）

当初、この学習者は「パスタ」を知らず、「サラダ」と聞き返しているが、その後、形を正しくマスターし、また意味についても学習している様子がうかがえる。

同様に、（6）は「耳をつねる」というコロケーションの学習が行われた場面である。実際にはテレビ電話なので、動作を通じて「つねる」の意味は正確に理解されていると言えよう。

（6）C：もし、じゅ、授業中、あー、先生の授業を、き、あー、聞かないで、ほかの人、ほかの人と、あー、しゃべるの、えー、のとき、えー、先生は、私、私たちの耳を、あー、しぼる？

J：耳を？　ぴゅ？
C：はい、●●です。
J：ああ。それはしないですかね。ふーん。
C：●。
J：しぼる。えーと、つー、つねる。
C：●あの、●●●●なんですか。あ、つねる。
J：うん。こうやって、こう、ぐっていうのね。　（04-05-3）

つまり、学習者と母語話者の心的辞書においては語彙の差があっても、わからない語が会話で出てくれば、必ず意味交渉が起こるため、結果的に実質語の使用語彙に差が見られなくなると考えられる。

他方、副詞に大きな差異が見られるのは、このような意味交渉が副詞では起こらないためであると考えられる。

6 おわりに

この論文では、真正性のある接触場面会話コーパスである『日中Skype会話コーパス』を紹介し、それを使った研究の端緒として、学習者と母語話者の語彙を比較した。結果、実質語には大きな差異は見られず、機能語、特に副詞に大きな差異が見られることが明らかになった。

実質語に差異が見られなかったことは、動詞や名詞がコミュニケーションの中心であることを意味している。これらの理解がなければ、コミュニケーションは成立しない。本コーパスは1時間以上というまとまった長さの会話のコーパスであるが、ある意味ではその長さの会話を成立させる前提条件として、動詞や名詞が話者間で共有されている必要があるとも言えるだろう。しかし、5節で確認したように、仮に学習者が知らない語があったとしても、接触場面での意味交渉を通じて、会話の中で語の学習が進むこともある。

他方、機能語は接触場面での学習はあまり期待できそうにない項目である。これらはコミュニケーションにおいては主役とは言えないかもしれないが、話者の日本語の自然さ、流暢さに大きく関わる項目である。そうすると、意味交渉の焦点とならないこれらの項目こそ、学習場面で重点的に扱うべきであるということも言えそうである。

ただし、最も大きな差異が見られた副詞は「単語のごみ箱」と呼ばれることもあるように、日本語学の研究としても日本語教育学の研究としても、遅れている分野である。この論文でも、マクロ視点による計量的な研究にとどまり、個々の副詞についてなぜ学習者が使えなかったり、あるいは多用したりするのかといった考察までは行えなかった。教科書で導入される副詞についても、いわゆる文法項目ほど十分な吟味が行われているとは言いがたい。これらの課題についてはまた稿を改めて取り組みたい。

最後に、『日中Skype会話コーパス』は従来の接触場面会話コーパスにはない種々の特徴を備えたコーパスであり、この論文ではそのほんの一端を分析したにすぎない。今後、多くの研究者の手で分析が行われていくことを期待する。

〈京都教育大学〉

謝辞
『日中Skype会話コーパス』の構築ならびに本研究の遂行には、JSPS科研費26770180の助成を受けた。

注
［注1］…… http://www.nakamata.info/database.htmlより。なお、本コーパスの詳細については同ページにリンクのある中俣（2015a）を参照のこと。
［注2］…… 北村・冨岡・川村（2009）はコーパスの出現文書数から語の難易度を求める試みであるが、「あさって」「おととい」のような語は基本語であるものの、コーパスに出現しにくいという問題点を指摘している。また、CSJとBCCWJの調整頻度レベルでは一番頻度が少ない曜日は木曜である（Tono, Yamazaki & Maekawa 2013）。
［注3］…… 詳細は、http://snowman.jnlp.org/ を参照のこと。
［注4］…… 語彙については学習者と母語話者の間よりも、話題の間に遥かに大きな違いがあることを中俣（2015b）は報告している。
［注5］…… 雪だるま独自の品詞体系による。この体系ではイ形容詞、ナ形容詞、連体詞が形容詞として分類される。なお、念の為にイ形容詞とナ形容詞を分けて分

析したところ、イ形容詞が0.917、ナ形容詞が1.195であった。イ形容詞の割合がやや低くなっているが、これは後述するように、学習者が「とても」を使うところで母語話者が「すごい」を多用していることが大きな原因である。

［注6］…… 副詞を機能語と見なすかどうかについては、いろいろな見解があるだろう。一つの目安となるのが山内（2012）の話題に従属するか否かという考え方である。実質語である名詞や動詞は話題に従属する。反対に機能語である助詞や接続詞は話題に従属しない。山内（2012）は副詞については言及していないが、時を表す副詞となる名詞（今日、去年など）は話題に従属しない名詞としており、同様の副詞も話題に従属しないと言えるだろう。少なくとも本研究で問題とした副詞は話題に従属しない。

参考文献

石川慎一郎（2012）『ベーシックコーパス言語学』ひつじ書房

宇佐美まゆみ・中俣尚己（2013）「「BTSJによる日本語話し言葉コーパス（トランスクリプト・音声）2011年版」の設計と特性について」第3回コーパス日本語学ワークショップ予稿集』pp.217–228.

北村達也・冨岡洋介・川村よし子（2009）「IDFを用いた単語レベル判定システムの構築と検証」『日本語教育方法研究会誌』16(1), pp.52–53.

田中牧郎・近藤明日子（2011）「教科書コーパス語彙表」『言語政策に役立つ、コーパスを用いた語彙表・漢字表等の作成と活用』pp.55–63.　2011文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21世紀の日本語研究の基盤整備」言語政策班

中俣尚己・漆田彩・小野真依子・北見友香・竹原英里（2013）「Skypeを活用した日中会話交流プログラム」『実践国文学』83, pp.132(25)–109(48).

中俣尚己（2015a）「日中Skype会話コーパスについて」http://nakamata.info/about_skype_corpus.pdf

中俣尚己（2015b）「「日中Skype会話コーパス」を用いた話題別語彙の抽出―「食」の場合」『第8回コーパス日本語学ワークショップ予稿集』pp.11–18.

橋内武（1999）『ディスコース　談話の織りなす世界』くろしお出版

山内博之（2012）「非母語話者の日本語コミュニケーション能力」野田尚史（編）『日本語教育のためのコミュニケーション研究』pp.125–144.　くろしお出版

Long, M. H. (1983) Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics, 4*(2), pp.126–141.

Tono, Y., Yamazaki, M., & Maekawa, K. (2013) *A Frequency Dictionary of Japanese*. London: Routledge.