エッセイライティングにおける語数の目安の指示の影響について

NICEのデータを見直してみる

杉浦正利

2016-05-07

学習者コーパス NICE は、2015年7月30日に、それまでのものとは別に新たに収集したNICE 3.0が公開されました。その後、さらに追加収集をし、NICE 3.1として近々公開予定です。今回の報告には、NICE 3.1 のデータも含めています。
NICE 3.0より前のデータは、NICE 2.2.2 として別になっています。
NICE 2.2.2 とともに公開されている説明文書にもすでに書いてあることですが、NICE 2.2.2 に収録されているデータは複数の種類のデータから成り立っています。
1時間で英文エッセイを書くというタスクですが、500語が目安であるということを指示した場合と指示していない場合とがありました。
また、監督者なしで執筆しそれをメールで送ってもらうという形で収集したものもありました。指示書には「本番（1時間）、作文（500語を目指してください）」とありました。英文エッセイの著作権譲渡の契約書には、著作物がどのようなものであるかの説明として「1時間で書けるだけ（もしくは500単語程度）」と説明してありました。
トピックについては、いずれのエッセイも11のトピックのうちの一つについて書いたものですが、その選び方は、すべて全く自由というわけではなく、トピックを指定して書いてもらった場合もありましたし、一人が二つ書いた場合は、二つ目を選ぶ際には、一つ目を除いた10から選ぶということもありました。

NICE 2.2.2 は、複数の由来のデータをひとまとめにしていましたが、こうした条件の違いが場合によっては何らかの影響を持つ可能性があります。

そこで、今回、データの由来ごとに、分けてそれぞれのデータの特徴について観察してみたいと思います。特に、500語という語数の目安が、どのように影響するか、という点を中心に見ていきたいと思います。

NICE2.2.2jp 342個のデータの内訳

	ファイル番号	ファイル数	由来	監督	指示の紙
NICE1jp	1～201	201	科研プロジェクト	有（ただしこのうち11個は監督なし・？）	有
	202～209	8	英語コーパス学会発表用追加分	無	無
NICE2jp	210～342	133	その後の追加分	有	無

NICE 2.2.2 の学習者データには、342個のファイルが含まれていますが、これは、大きく3つに分けられます。

1） JPN001からJPN201：当初の科研プロジェクトで収集。これを便宜上「NICE1jp」と呼ぶことにします。

これらのデータ収集において、確認したところ、11個のファイルについて、監督者なし、もしくは、不明のものがありました（約5%）。

監督者ありの場合に使用した「指示の紙」には、英文エッセイの目安は500語であるということが書かれていました。

ゆえに、NICE1jpの約95%は、監督あり、500語目安の指示あり、という条件と言えます。

2） JPN202からJPN209：英語コーパス学会第29回大会での研究発表「英語学習者コーパスにおける作文テーマの影響」のために追加収集。

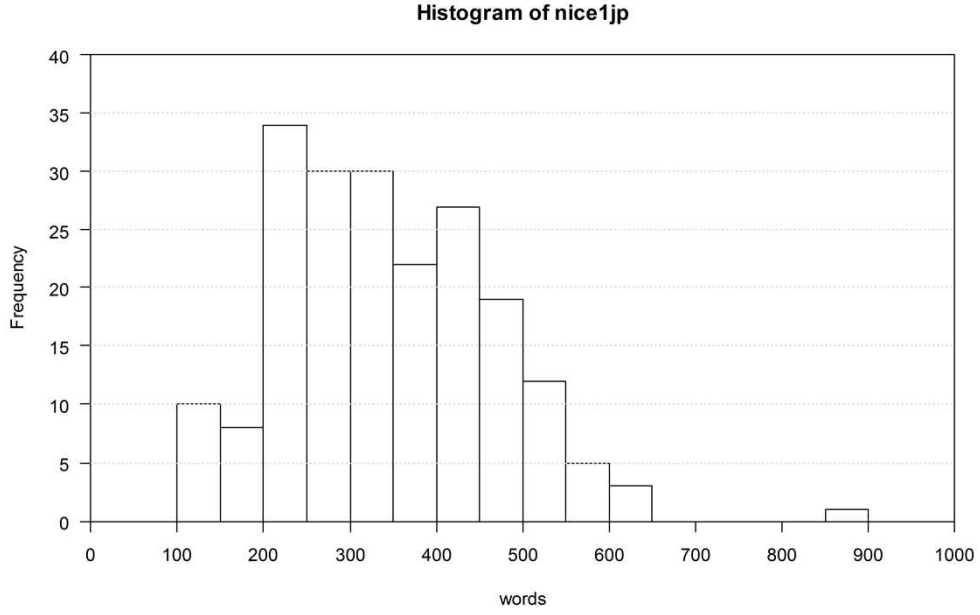
これらは、学会発表のために追加で収集したデータでした。メールもしくは口頭で依頼をし、メールの添付ファイルでデータを送ってもらいました。

ですので、これらは監督者がなく、指示の紙もなかったので、今回は分析対象から外します。

3） JPN210からJPN342：その後、別の研究のためにトピックをmoneyとschool educationに限って追加収集。これを便宜上「NICE2jp」と呼ぶことにします。

これらは、監督はすべてありましたが、500語目安の紙はなかったので、NICE2jpは、監督あり、500語目安の指示なし、という条件と言えます。

NICE1jpの201個のファイル



NICE1jpの201個のファイルの平均語数は、342語でした。
500語以上は21個、500語未満が180個（約90%）でした。

3種類のNICEのデータ

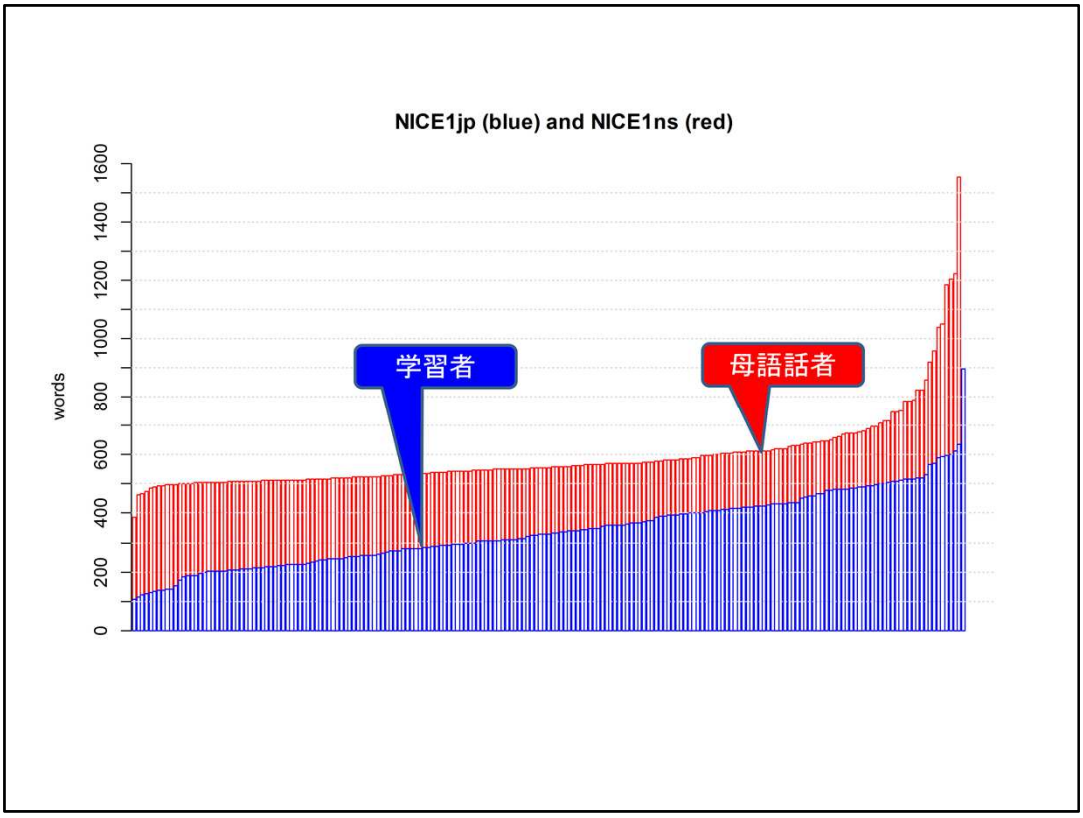
	JP (ファイル数)	NS (ファイル数)
NICE1*（科研分）	NICE1jp (201)	NICE1ns (200)
NICE2（追加分）	NICE2jp (133)	NICE2ns (10)
NICE3（新規分）	NICE3jp (185)	NICE3ns (36)

*NICE1については、JPN202からJPN209までの8ファイルを除外する。

NICE1jpは、当初の科研で収集し、監督あり、500語目安の指示ありの条件、
NICE2jpは、追加分で、監督あり、500語目安の指示なしの条件です。

これらとは別に、新たに収集したデータを、NICE3jpと呼び、比較します。
NICE3 は、すべて、監督あり、500語目安の指示なしの条件です。

そして、それぞれの時期に集めた母語話者データを、それぞれ、NICE1ns、NICE2ns、
NICE3nsと呼ぶことにします。

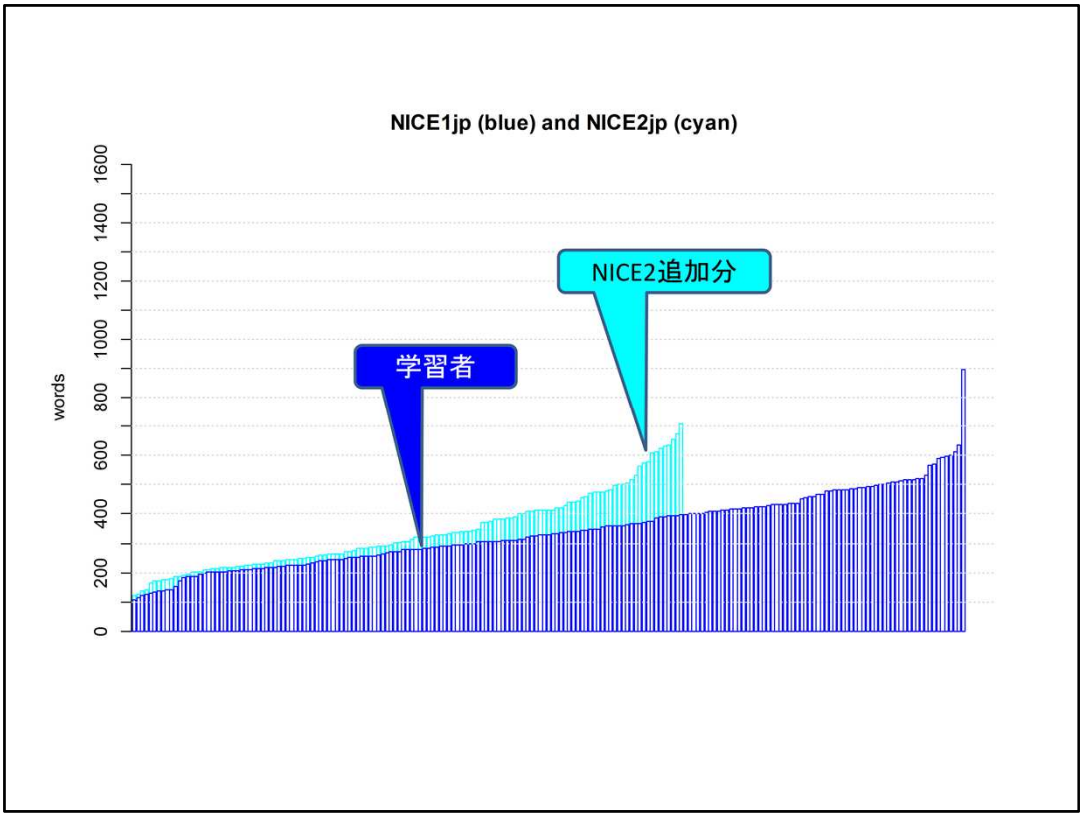


まず最初に、当初の科研で収集したデータであるNICE1jpとNICE1nsデータの各ファイルを総語数順に並べ替えてみます。

青い学習者のデータ NICE1jp は徐々に右肩上がりになっています。

これに対し、赤い母語話者データ NICE1ns は、500語から600語のあたりでかなり平らになっています。母語話者には、1時間で英文エッセイを書くようにと指示をした場合に、目安が500語であると示したことが影響し、結果的に500語くらいで英文エッセイをまとめるという課題になってしまったと考えられます。
同様に500語が目安であると言われても、学習者の場合は、そもそも1時間で500語の英文エッセイを書くこと自体が難しかったと思われます。500語以上のファイルは21個で、全体の約10%でした。

このことから、1時間で500語を目安に英文エッセイを書く、という同じ指示であっても、学習者と母語話者とでは、その指示の持つ意味が、結果的に違ってしまったと考えられます。指示の際の語数の目安がこのような影響を及ぼすとは当時考えが至りませんでした。



追加で収集した NICE2jp のデータを同様に、総語数で並び替えて水色で NICE1jp のグラフに重ねてみます。
NICE2jpの水色のグラフは、若干、右肩上がりの傾きが急に見えますが、NICE2jpはファイル数が133個で、NICE1jpの201個よりも少ないので、急になっているように見えると言えます。NICE2jpで、500語以上の総語数のファイルは16個で、全体の約12%です。
NICE1jpでの500語以上のファイル数が約10%ですから、NICE2jpの方がおよそ2%分だけ500語以上のファイル数が多いことになりますが、全体の傾向としてはほぼ同じと考えらえると思います。

NICE2jpでは、すべて監督ありですが500語目安の指示はありませんでした。
NICE1jpでは、500語目安の指示がありました。

ゆえに、学習者が、1時間で英文エッセイを書く場合、500語目安の指示があってもなくても、結果的には、500語を超えて書くことは9割近くの人にとって困難なので、ほとんど影響しないと考えられます。

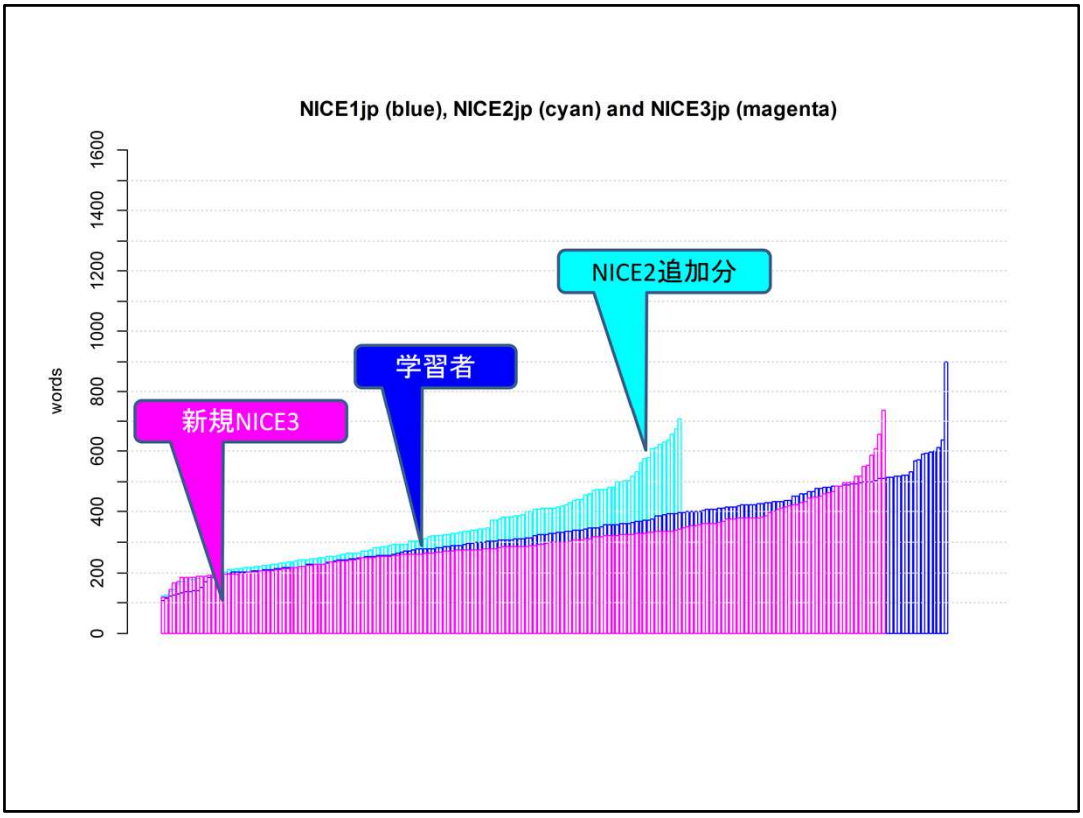
NICE1jpとNICE2jpとの差の検定

- ウィルコクソンの順位和検定

- W = 14130.5, p-value = 0.3768
- 効果量 r = 0.0484

- p値が有意水準5%を超えており、検定の結果、二群の代表値に差があるとは言えない。

また、NICE1jpとNICE2jpとのデータに、差があるかどうかを統計的に確かめるために、ウィルコクソンの順位和検定を行ってみました。
その結果、p値は0.3768となり、有意水準 0.05 より大きく、二群の代表値に差があるとは言えないという結果になりました。



さらに、新たに収集した NICE3jp のデータを重ねてみます。ファイル数が185個とNICE1jpよりも少ないですが、右肩上がりの傾向はNICE1jpやNICE2jpとほぼ同じであるように見えます。

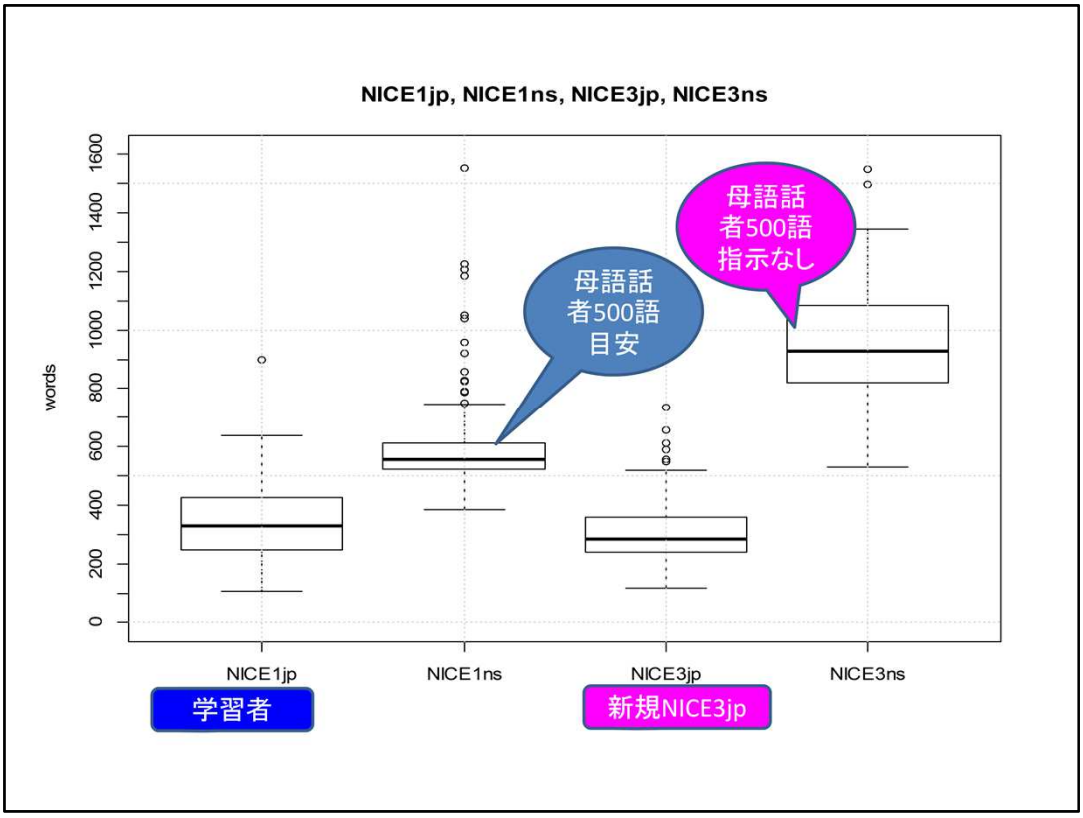
NICE3jp で総語数が500語を超えたファイルは9つで、全体の約5%でした。
NICE2jp と同様に、監督ありで500語目安の指示がないのに、500語を超えるファイル数の割合が少なくなっています。

これは、おそらく学習者の中に含まれる大学院生の割合が影響しているのではないかと思われます。

学部生と大学院生

	学部生	大学院生	不明
NICE1jp	145	53 (26%)	3
NICE2jp	97	36 (27%)	0
NICE3jp	179	6 (3%)	0

それぞれ3種類のデータに含まれる学部生と院生との割合を見てみると、NICE1jp と NICE2jp では、大学院生がおよそ26%と27%であるのに対し、NICE3jp は185人中6人のおよそ3%しか大学院生がいません。



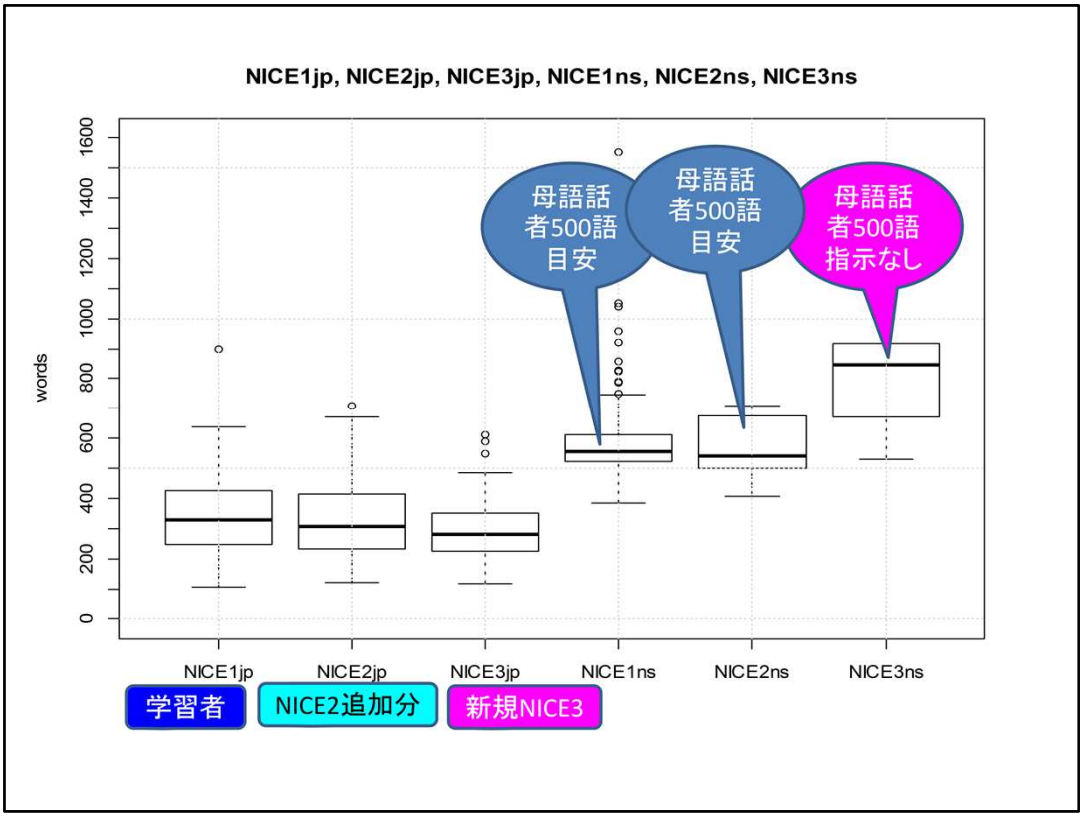
データの分布の様子を別の角度から箱ひげ図で観察してみます。上下の真ん中の四角の箱の部分にデータ全体の50%が入るという図です。まん中の太い横線は中央値です。

4つのグラフのうち、左側の二つが当初の科研で収集したデータ NICE1jp と NICE1ns です。

母語話者データである NICE1ns が、500語から600語に集まっていることがわかります。

これに対し、右側の二つは、新規に収集したNICE3のデータです。学習者・母語話者とも、監督ありで、500語目安の指示はない、という条件です。この NICE3 の場合、学習者データは、 NICE1jp に近い分布をしていますが、少し少なめであることがわかります。これは、NICE1jp には大学院生が約26%含まれているのに対し、NICE3jpの方は大学院生は約3%であったことに由来しているのではないかと思われます。

NICE1nsの母語話者データに比べ、NICE3ns の母語話者データは、36個と少ないですが、データの分布としては、およそ800語から1100語あたりに分布しています。これは、 NICE1ns の分布とはずいぶん違うと言えます。これは英文エッセイを書く際に目安として500語という語数があったかなかったかによる影響ではないかと考えられます。



3種類の学習者データと、3種類の母語話者データの分布を見てみます。

左側3つの学習者データに関しては、NICE3jp で少し語数が少ないですが、だいたい3種類とも同じような分布をしていることがわかります。

これに対し、右側3つの母語話者データに関しては、先ほどみたように、NICE1ns と NICE3ns とではずいぶん分布が違い、その原因は500語目安の指示があったかどうかに起因すると思われます。

NICE2ns は、ファイル数は10個と少ないですが、この収集の際には、500語の目安を示しています。NICE2ns の分布は、同様に500語の目安を示した NICE1ns の分布に近いといえるのではないでしょうか。

500語目安の影響

- 学習者の場合、500語目安の指示の有無が強く影響するとはいえない。
  - 多くの学習者にとって1時間で500語のエッセイを書くことは難しいと思われる。
- 母語話者の場合、目安の指示がある場合、強く影響を受ける。
  - 指示がないと、過半数の人が、1時間で約800～1100語のエッセイを書く。
  - 指示があると、500語程度でエッセイをまとめようとすると思われる。
- 「1時間で500語を目安に」という同じ指示が、学習者と母語話者とで違う影響を与えたと考えられる。
- その影響を考えなかったことはデータ収集時に配慮が足りなかった。

以上、NICE のデータを、その由来ごとに種類を分け、比較してみることで、英文エッセイを書く際の指示に500語という目安があることが、学習者と母語話者とで、違う影響を与えていることが分かったといえます。

「1時間で英文エッセイを書く」という指示をする場合に、500語の目安を示しても、学習者の場合は、およそ9割ほどは500語を超える英文エッセイを書くことはなかったのに対し、母語話者の場合は、500語の目安があれば500語を実際に目安として英文エッセイを書き、その目安がない場合はおよそ800語から1100語くらいの英文エッセイを書いています。同じ条件といっても「1時間で500語を目安に英文エッセイを書く」という場合と、「1時間で英文エッセイを書く」という場合とで、500語の目安の有り無しが、学習者には強い影響は与えないとしても、母語話者の場合には強く影響するといえるでしょう。ただし、学習者の場合でも、目安として示す語数が、例えば300語とか、もっと少なかったら、語数の目安の指示がもっと影響を与えていたのではないかと推測されます。

当初の科研でデータ収集をする際に、この500語を目安とするということが、学習者と母語話者の英文エッセイライティングにおいて、これだけ違う影響を与えるということは考えていませんでした。配慮が足りなかったことを反省しています。