Vol.2014-DBS-159 No.23
Vol.2014-IFAT-115 No.23
2014/8/2

老いと＜ことば＞：ブログ・テキストから測る老化

荒牧英治[†1 †3]　久保圭[†2]　四方朱子[†1]

言語能力は人生における経験の結晶であり，加齢によって損なわれることがないとされる．しかし，その一方で，文法能力など，一部の能力の加齢による低下が示されることもある．このように，老化と言語能力の関係については不明な点が多い．この原因は，次の 2 つによるところが大きい．まず，研究対象である高齢者から大規模なデータを得るのが困難であること．次に，言語はさまざまな能力の総体であり，調査ごとに測っている言語能力が異なることである．そこで，本研究では，Web 上の文章を利用する．まず，50 代から 80 代の高齢者や小中学生，第二言語習得者，認知症患者のブログや作文を集めた．また，測定に関しては，語彙に関するものや構文に関するものなど，さまざまな指標を用いた．この結果，高齢者は，使う言葉の種類が減る可能性があること，さらに，難易度の高い言葉から使用頻度が減ることが明らかになった．この知見を応用することによって，老化や認知症の早期発見の可能性があり，今後の応用が期待される．

Aging and Language: Blog Texts as an Aging Index

Eiji ARAMAKI[†1 †3]　Kay KUBO[†2]　Shuko SHIKATA[†1]

Preceding study claims that one's language abilities develop over long period of time and improve with age. On the other hand, some study reports that some parts of language abilities, such as grammatical ability, show some decrease in elder people. Since one's language ability is often shown as the aggregation of multiple human abilities, it is difficult to solely extract his/her language ability out of his/her written texts. This study, thus, analyzes texts by using multiple linguistic measures. The corpora cover school students (children attending primary to junior high school, age 6 to 15 years old), elders (age 50 to over 80 years old), Japanese as the Second Language learners, and a dementia patient (Alzheimer type). As a result, this study shows that the lexical richness decreases, and difficult vocabularies tend to be especially lost from elders. This study also displays the possibility of detecting dementia in its early stage.

1. はじめに

近年，日本におけるブログ文化の定着に伴い，パソコンを用いた高齢者による言語表現の機会が顕著に増加してきている．推論能力や短期記憶など，老化によって衰退する能力とは異なり，言語能力は言語理解や語彙，一般的事項の情報量や常識などの総合的な能力である．この点において，言語能力は，老化によって低下するといわれる，情報を素早く処理するような流動性の能力とは区別され，「結晶化知性 （crystallized intelligence）」 [1, 2] とよばれることもある．

言語能力は長期におよぶ学習や経験によって発達するものであり，一定レベルまで発達した後は，加齢によっても衰えにくいとされる[3]．事実，高齢者の文章の創造性について注目した解説もある[4]．その一方で，構文をあやつる能力は，70 代後半を境に低下しはじめるという報告もある[5]．このように，老化と言葉との関係については，矛盾を含む結果がある．この原因としては，次の 2 つが考えられる．まず，研究対象である高齢者から大規模なデータを得るのが困難であること．次に，言語はさまざまな能力の総体であり，調査ごとに測っている言語能力が異なっているということである．

より緻密な結果を得るため，大規模なコホート研究によって，数十年の言語能力の経過を観察する試みが行われている．その結果，老化や認知症などとさまざまな加齢による能力との関係は徐々に明らかになりつつある[6, 7]．しかし，これには大変な労力を必要とする．

そこで，我々は，自主的に執筆されている大量のブログを利用することで，言語能力の調査が可能ではないかと考えた．ブログは本邦において広く普及しており，実に世界の 37%のブログ投稿が日本語によるものである[8]．このブログ執筆者の中には，高齢者も多く含まれる．このデータを用いることで，高齢者の文章の特徴をとらえることができる可能性がある．

本研究では，高齢者をはじめとして，小中学生，第二言語として日本語を学習している学生，認知症患者など，さまざまな執筆層からなるデータを集め，言語能力の計測を行なった．

†1 京都大学
Kyoto University
†2 大阪大学
Osaka University
†3 科学技術振興機構 さきがけ
JST PRESTO

Vol.2014-DBS-159 No.23
Vol.2014-IFAT-115 No.23
2014/8/2

結果,高齢期には,使う言葉の種類が減る可能性があり,さらに，より難易度の高い言葉から使用が減ってゆくことが示唆された．この知見を用いることによって，老化や認知症の早期発見の可能性があり,今後の応用が期待される．

2. 材料

本研究で用いる材料データは，以下のとおりである．これらデータの構成を表 1 に示す．

- ***L1***: 小中学生作文データ
- ***L2***: 日本語学習者作文データ
- ***E:*** 高齢者ブログデータ
- ***ΔE***: 高齢者長期ブログデータ
- ***ΔD***: 認知症患者長期ブログデータ

2.1 L1: 小中学生作文データ

小中学生作文データは，郵便事業株式会社が主催する「第 42 回手紙作文コンクール」の手紙作文部門における入賞作品の全テキストを収集したものである．小学生 97 名（低学年男子 22 名，低学年女子 26 名，高学年男子 17 名，高学年女子 32 名）および，中学生 48 名（男子 5 名，女子 43 名）の作文データを用いた．1 文章あたりの平均文数は 30.9 文である．

2.2 L2: 日本語学習者作文データ

日本語学習者作文コーパス[a]は日本語を学ぶ留学生のテキストである．このコーパスでは，学習レベル別に 304 名ぶんの作文テキストデータが公開されている．本研究では，初級と上級の作文(初級 31 名，上級 124 名)を用いた．1 文章あたりの平均文数は 17.0 文である．

2.3 E: 高齢者ブログデータ

高齢者ブログデータはブログ・リンク集を用いて無作為に抽出した高齢者のブログである．4 つのリンク集[b]から 50 代・60 代・70 代・80 代以上について，それぞれ男女別に 10 名ずつ無作為に選択し，タイトル抜き，本文のみ，他者引用は避けるという 3 つの基準のもとそれぞれ 500 文を抽出した. ただし, 80 代以上十分な数が集まらず, 男性 9 名，女性 8 名を収集した．

2.4 ΔE: 高齢者長期ブログデータ（5 年執筆; n=5）

上記高齢者ブログデータ(E)のうち，収集時に 70 歳以上，かつ，5 年以上の期間にわたって執筆している男性を，高齢者長期ブログデータ(ΔE)として区分し，執筆年ごとに 500 文ずつを収集した(n=5).

2.5 ΔD: 認知症患者長期ブログデータ(5 年執筆; n=1)

認知症患者ブログデータは，ある認知症患者（最終執筆時 70 代男性）が綴ったブログを収集したものである．このブログの執筆期間は約 6 年間であり，執筆開始から 5 年間のテキストを執筆年ごとに 1200 文ずつ収集した．

表 1：データ構成　※カッコ内はデータ数（＝人数）

L1	小学低学年 男性（22 名） 女性（26 名）	小学高学年 男性（17 名） 女性（32 名）	中学 男性（5 名） 女性（43 名）	
L2	初級 （31 名）	上級 （124 名）		
E	50 代 男性（15 名） 女性（15 名）	60 代 男性（15 名） 女性（15 名）	70 代 男性（15 名） 女性（15 名）	80 代以上 男性（9 名） 女性（8 名）
ΔE	Year 1 男性（5 名）	Year 2	Year 3	Year 4	Year 5
ΔD	Year 1 男性（1 名）	Year 2	Year 3	Year 4	Year 5

表 2: 本研究で用いる言語能力に関する指標．†は筆者らのグループが開発した指標

指標	略記	説明	単位／方法	対象
D-LEVEL 日本語版†	LEV	文の複雑さ示す	単文 経験則	文法 能力
構文木の深さ	DEP	文の複雑さ示す	単文 統計量	文法 能力
日本語学習辞書レベル	JEL	語彙の難しさ	単語 経験則	語彙 能力
特殊性†（逸脱率）	FPU	語彙の特殊性	単語 統計量	語彙 能力
具体性	NER	固有名詞の割合	単文 経験則	語彙 能力
タイプ・トークン割合	TTR	語彙の量	文章 統計量	語彙 能力
機能表現レベル	FNC	難易度	単文 経験則	その他
ポライトネス	PLT	丁寧さ	単文 経験則	その他

3. 方法

本研究で調査する言語能力は，これまでに開発してきた指標，また，本研究において初めて導入された指標を含めて多数ある．各指標は，以下の 3 つの観点で分類できる．

- **単位**：語単位で算出され，それを全文で平均するタイプのもの（単語指標）や，文単位で算出され，それを全文で平均するタイプのもの（単文指標），使用語彙数など，全文から導出されるタイプのもの（文章指標）がある．
- **方法**: 経験則によりレベル付けされたもの（経験則）や, 大規模コーパスから統計的に算出されるもの（統計）がある．
- **対象**：文法能力に注目したものや語彙能力に注目したものなど，測定対象による分類が可能である．

これらの観点から指標を分類した結果を表 2 に示す．以下に，各手法の定義を述べる．

- **D-LEVEL 日本語版（LEV）**: 文の複雑さを示す．

a http://sakubun.jpn.org

b http://www.tobyo.jp/
http://ameblo.jp/ryu280/
http://mooyan.air-nifty.com/
http://blog.livedoor.jp/mataichi000/

Developmental level [9] によって 7 つのレベルに分類された英語の構文を，それに相当する日本語の構文にほぼ置き変えたものである．詳細については，ウェブページを参照[c]．なお，これは文ごとに算出し，平均した．

- **構文木の深さ（Mean of Tree Depth; DEP）**：文の複雑さを示す．係り受けの観点から算出される，構文木の最大の深さ．日本語の構文木は，係り受け解析器 KNP[d]を用いて得た．文ごとに算出し，平均した．
- **頻度・使用者数比（Frequency per User Popularity; FPU）**：語の特殊性を示す．語の特殊性は，語の出現頻度／語のユーザ数とした．この値が低いほど，その語は一般的であり，高いほど，ユーザ数が出現頻度と比較し，少ない語であることを示す．スラングや専門用語などは高い値を持つ．語のユーザ数は，ソーシャルメディア上の 10 万人の発言を 8 ヶ月間調査して得た．詳細は文献 [10] を参照．語ごとに算出し，全単語の平均を算出した．
- **日本語学習語彙レベル（Japanese Educational Lexicon Level; JEL）**：語彙の難易度を示す．難易度は日本語学習辞書[e]に収載されている語彙レベルを用いた．語彙レベルは，1（初級前半），2（初級後半），3（中級前半），4（中級後半），5（上級前半），6（上級後半）に分けられる．詳細は文献 [11] を参照．語ごとに算出し，全単語の平均を得た．
- **TYPE・TOKEN 割合（Type Token Ratio; TTR）**：Type（異なり語数）と Token（延べ語数）の比率（Type／Token）．この値が大きいほど，語彙量が多い．文章全体で集計した．
- **機能表現難度（Difficulty of Functional Expression; FNC）**：機能表現の難易度を示す．この値が大きいほど，文章内で用いられている機能表現の難易度が高い．難易度の定義は「日本語機能表現辞書つつじ」[12] で設定されている難易度による．難易度は A1, A2, B, C, F の 5 段階に分かれており，これを 1 (A1) から 5 (F) に変換した．文ごとに算出し，平均した．
- **ポライトネス（Politeness of Functional Expression; PLT）**：機能表現のポライトネスの度合いを示す．この値が大きいとき，機能表現が丁寧であることをあらわす．ポライトネスは「日本語機能表現辞書つつじ」[12, 13]の分類を採用した．分類では機能表現が常体 (normal)，敬体 (polite)，口語体 (colloquial)，堅い文体 (stiff) の 4 種類に分けられており，口語体 (colloquial)＝1，常体 (normal)＝3，敬体 (polite)＝5，堅い文体 (stiff)＝5 に変換した．文ごとに算出し，平均した．
- **具体性（Named Entity Ratio; NER）**：固有名詞の割合を示す（固有名形態素数／全名詞形態素数）．この値が大きければ，文章の内容がより具体的であることを示す．固有名詞の判定は，形態素解析器 JUMAN [14] の解析結果による．地名，数詞，固有名詞の割合を文ごとに算出し，平均した．

c http://mednlp.jp/dlevel
d http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
e http://jishokaken.sakura.ne.jp/DB/

4. 関連研究

4.1 言語能力とは

一般に，言語能力は「話す (speech)・聞く (listening)・読む (reading)・書く (writing)」の 4 つに大別される．本研究はウェブの発達によって大量のデータが集積されつつある＜書かれた＞データに注目する．

この＜書く＞能力は，語彙に関する能力（以降，語彙能力とよぶ）と文法に関する能力（以降，文法能力とよぶ）の 2 つに大別される [15][16]．本研究では，これら 2 つの能力を，さらに細かく分化する．まず，語彙能力については，語彙の難易度と使用語彙量という 2 つの指標から調査をおこなう．また，文法能力については，構文木の深さや D-level を測定する．このような多岐に渡る指標を用いて日本語文章の調査をおこなった研究は，管見の限り存在しない．

4.2 疾患と言語能力

失語症などの言語との関連が深い疾患を除けば，罹患中の言語能力の変化についての調査は少ない．しかし，85 歳以上の 40%が罹患している認知症[17]については近年研究が進んでいる．Snowdon は，認知症の症状が認められ始める 50 年前から語彙能力が低いと報告した[7]．また，Kemper によると，英文における認知症の進行度は構文能力とも相関関係にあり，症状が進行するにつれ，構文能力の顕著な低下がみられるという[1].

ただし，ある人間が認知症か否かを判定するための項目に，言語能力に関するものがあることを考慮すれば，認知症患者が言語能力の低下を示すことは，ある種のトートロジーであり，言語能力と認知症，さらに老化の関係は，複雑な様相を呈しているといえる．

本研究で試みたように，文章から認知症などの疾患を推定することが可能となれば，測定の侵襲度が低く，かつ，超早期からの検出が可能となる．

4.3 老いと言語能力

前述のとおり，一般に，老化と言語能力とは相関しないといわれている．例えば，一部の言語能力は，高齢者においても伸び続けるため，老人が理解している語彙数は，若者の約 1.3 倍だという報告もある[18]．その一方で，人間が構文をあやつる能力は，70 代後半を境に低下しはじめるとい

Vol.2014-DBS-159 No.23
Vol.2014-IFAT-115 No.23
2014/8/2

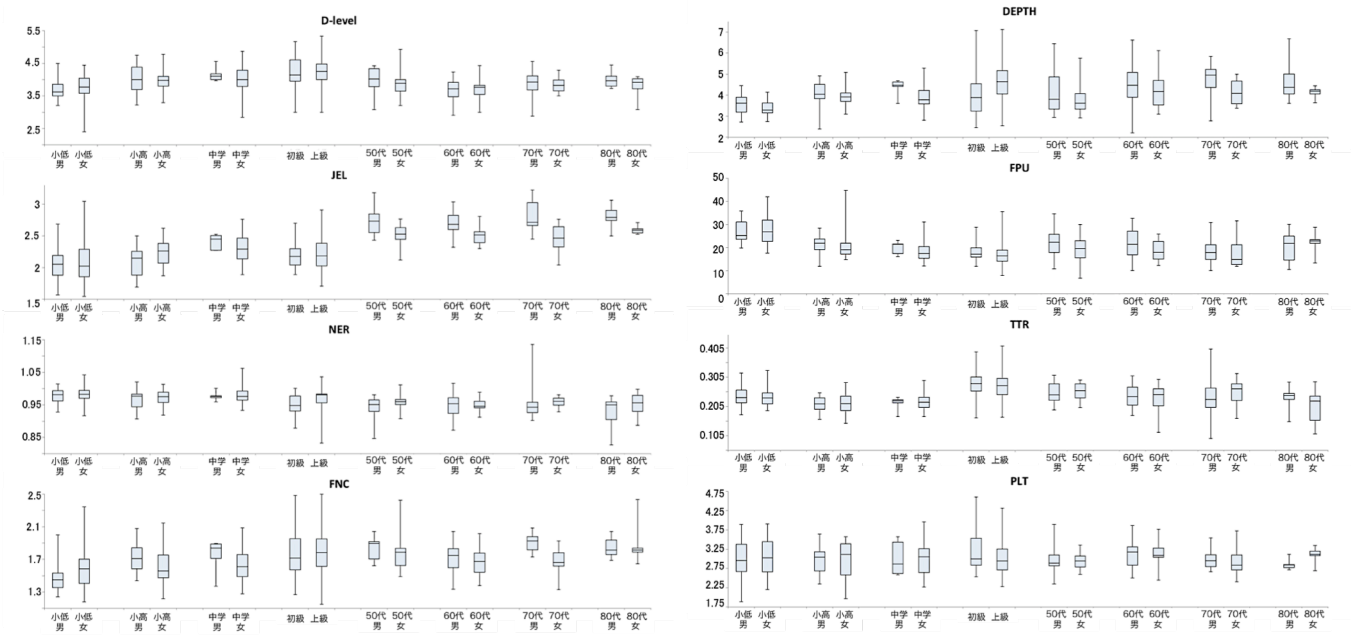

図 1: コーパスと指標.

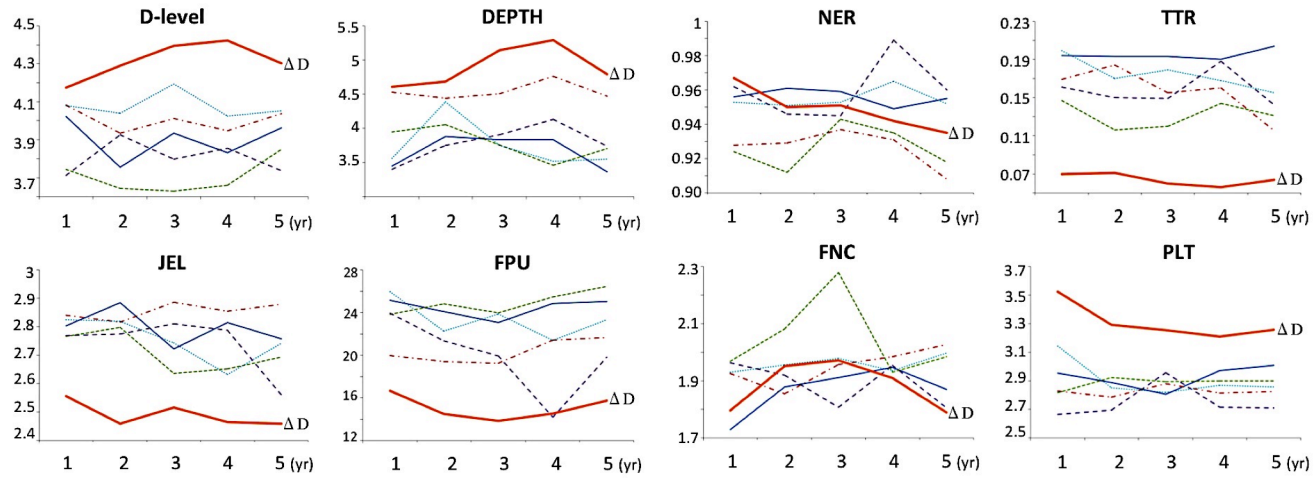

図 2: ΔE および ΔD における指標の変化.

う報告もある[5]．このように，老化と言語能力については不明な点が多い．本研究は，これを解明する一助となることを目指す．

5. 結果と考察

結果を図 1 と図 2 に示す．図 1 にコーパス（L1，L2，E）ごとの指標を箱ひげ図で示す. 図 2 に長期ブログデータ（ΔE と ΔD）の年ごとの指標の変化を示す.

5.1 小中学生と高齢期（L1 と E）

小中学生期（L1）では多くの値が変化する．文法能力（LEV と DEP）と語彙能力（JEL）が上昇し，特殊性（FPU）が減少する．この時期に言語能力が上昇していると考えられるため，各尺度がある程度，言語能力の成長を捉えることに成功していることが分かる．

一方，高齢期（E）においては，年齢とともに変化する値は少ない．変化が見られるのは，文法能力（DEP）と語彙能力（JEL）であり，いずれも年齢とともに高くなっている．このことから，次の 2 つの可能性が示唆される．（仮説 1）高齢者においても言語能力の一部の発達は継続する．あるいは，（仮説 2）元来言語能力の高い人間のみが高齢者になっても執筆を行なっている（選択バイアスによる見せかけの上昇）.

5.2 高齢期における個人の変化（ΔE）

以上の仮説のもと，次に，高齢期のおける個人の変化（ΔE）を観測する．このデータでは，5 名の人物それぞれの 5 年の経過を観測しているため，選択バイアスはないと考えられる.

その結果，高齢者の年代別データ（E）で見られた能力

の上昇は見られず，逆に語彙能力（JEL と TTR）の減少を認めた．このことから，高齢者年代別でみられた能力の上昇は，仮説 2 のとおり，選択バイアスによる見せかけの上昇である可能性が高いと考えられる．

一般に高齢者において，言語能力は衰退しないとされているが，本結果からは，語彙能力以外の能力は高齢者においても低下を認めず，部分的に過去の報告を支持していると言えよう．

5.3 第一言語学習と第二言語学習（L1 と L2）

前述したように，小中学生においては，年齢の上昇とともに多くの言語能力の変化が確認された（DEP，LEV，FNC，JEL，FPU）．これらは第一言語学習時の変化と考えることができる．

しかし一方で，第二言語学習者においては，初級と上級の間での変化はわずかしか見られず，唯一，文法能力（DEP）のみの上昇が確認された．このことから，第一言語学習時と第二言語学習時においては，言語能力の変化に違いがある可能性が示唆された．

小中学生が L1（つまり母語）を習得する際には，それまでに習得した個々の単語の音韻的・統語的・書記的関係を随時発見し，加齢とともに言語能力をある期間まで急加速度的に発達させていく．その一方で，成人学習者が L2 を習得する際には，すでに習得している L1 の語彙体系を随時借用することが多いため，新たな L2 の語彙の意味は，その形式と 1 対 1 で習得される．そのため，初級と上級を比較しても，語彙の多様性に差がない可能性があり，本結果は，これを裏付けているといえる．

5.4 男女差

男女別に集計可能なデータ（L1 と E）を用いて，性差をみる．

小中学生において，一部の能力（DEP，FNC と JEL）に性差が認められ，いずれも最終的には男性が高値を示している．文法能力（DEP）は小学生時には性差がなく，中学時に性差が顕著となる．語彙能力（JEL）は，小学高学年までは女性が高いが，中学時からは男性が高値となる．機能表現レベル（FNC）は，小学低学年時には女性が高値であるが，高学年時から逆転し，男性が高値となる．

高齢期においても性差が認められるのは，小中学生時（L1）と同様の指標である（DEP，FNC と JEL）．これらは小学校高学年や中学などの L1 後期時と同様に，男性の高値を示している．

これらの結果から，男性は女性よりも複雑な表現を使用し（FNC），文が複雑になりがちなこと（DEP），さらに，用いる単語も難解な傾向がある（JEL）ことを示している．ただし，これが，ただちに男性の言語能力の高さを意味するものではないことに注意されたい．

5.5 老化における変化とは

老化において次の 2 点が起こることが示唆された．まず，高齢期（ΔE）には，使う言葉の種類が減る可能性が示唆された（TTR の減少，$R^2 = 0.47$）．さらに，難易度の高い言葉から使用頻度が減っていることも示唆された（JEL の減少，$R^2 = 0.83$）．

5.6 認知症における変化とは

認知症を発症すると次のことが起こると考えられる．第一に，用いる言葉の種類が減り，時間経過と共にさらに減少する（TTR 減少）．第二に，難易度の高い言葉も使用頻度が減少する（JEL 減少）．そして，文の構造が複雑化し，言葉づかいが丁寧になる．

また，認知症患者と健常高齢者を比較すると，いくつかの指標（TTR および JEL）に大きな差がある．このことから，認知症のごく初期の段階において，言語能力（TTR および JEL）の甚大な低下が起こる可能性がある．これらの知見を用いることで，老化や認知症の早期発見が可能になるかもしれず，今後の応用が期待される．

ただし，認知症例に関してはサンプルが 1 例のみであり，今後，大規模な検討が望まれる．

5.7 本研究の限界と今後の課題

本研究にはいくつかの限界がある．最も大きな問題は，データ選択バイアスであろう．小中学生時（L1）のデータとして，作文コンクールの入賞者のデータを用いているが，これは，L1 年代の最も言語能力の高い群を抽出していることに相当すると考えられる．このため，平均的な発達の過程の議論をすることは困難である．

また，高齢期（E）においては，ブログ執筆者のデータを用いている．ここではブログを執筆している群を抽出することが選択バイアスとなり得る．特に，70 代後半や，または，80 代以上の後期高齢者になるほど，ブログを執筆する（あるいは執筆できる）ということは，ある種の特殊な群であり，ここでもバイアスが生じる．

長期執筆群（ΔE）についても，長期の執筆に耐えうる群のみが選択されていることとなり，そこには大きな選択バイアスがある．また，執筆を続けることによる執筆能力の上昇も予想され，問題を複雑化する．

以上のような多くの問題が考えられるため，現在，より統制のとれたサンプリングを計画している．

6. おわりに

老化と言語能力の関係を調査するのは難しい．これは言語がさまざまな能力の総体であること，大規模な調査対象が得られにくいことなどが要因である．そこで，本研究では，高齢者や小学生や中学生，第二言語習得者，認知症患者のブログや作文を集め，さまざまな言語能力を測定した．この結果，高齢者は，使用する言葉の種類が減る可能性があること（TTR 低下），さらに，難易度の高い言葉から使用頻度が減ること（JEL 低下）が示唆された．

高齢者におけるこれらの言語能力と，認知症患者のそれ

と比べると，大きな言語能力（特に TTR および JEL）の差が認められた．この知見を用いることで，老化や認知症の早期発見が可能になるかもしれず，今後の応用が期待される．

謝辞 本研究は JST さきがけ「自然言語処理による診断支援技術の開発」プロジェクトの助成を受けた．

参考文献

1. Cattell, R.B., *Abilities: Their structure, growth, and action*. New York: Houghton Mifflin. 1971.
2. Horn, J.L. and R.B. Cattell, *Age differences in fluid and crystallized intelligence.* Acta Psychologica, 1967. **26**.
3. Hampshire, A., et al., *Fractionating human intelligence.* Neuron, 2012. **76**(6): p. 1225-37.
4. Adams-Price, C.E., *Aging, Writing, and Creativity*, in *Creativity & Successful Aging*. 1998, Springer. p. 269-287.
5. Kemper, S., J. Marquis, and M. Thompson, *Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content.* Psychol Aging, 2001. **16**(4): p. 600-14.
6. Kubo, M., et al., *Trends in the incidence, mortality, and survival rate of cardiovascular disease in a Japanese community: the Hisayama study.* Stroke, 2003. **34**(10): p. 2349-54.
7. Snowdon, D.A., et al., *Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study.* JAMA, 1996. **275**(7): p. 528-32.
8. ComScore. *Worldwide Internet Audience has Grown 10 Percent in Last Year, According to comScore Networks*. Available from: http://www.comscore.com/Insights/Press_Releases/2007/03/Worldwide_Internet_Growth.
9. Cheunga, H. and S. Kemper, *Competing complexity metrics and adults' production of complex sentences.* Applied Psycholinguistics, 1992. **13**(1): p. 53-76.
10. Aramaki, E., et al. *A Word in a Dictionary is used by Numerous Users*. in *International Joint Conference on Natural Language Processing (IJCNLP2013)*. 2013. Japan.
11. 砂川有里子, *学習辞書編集支援データベース作成について -『学習辞書科研』プロジェクトの紹介」*. 日本語教育連絡会議論文集, 2012. **24**.
12. 松吉俊, 佐藤理史, and 宇津呂武仁, *日本語機能表現辞書の編纂.* 自然言語処理, 2007. **14**(5): p. 123-146.
13. 松吉俊 and 佐藤理史, *文体と難易度を制御可能な日本語機能表現の言い換え.* 自然言語処理, 2008. **15**(2): p. 75-99.
14. Kurohashi, D.K.a.S. *A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis*. in *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2006)*. 2006.
15. Kintsch, W. and J. Keenan, *Reading rate and retention as a function of the number of the propositions in the base structure of sentences.* Cog Psych, 1973. **275**: p. 528-532.
16. Turner, A. and E. Greene, *The Construction and Use of a Propositional Text Base*, in *Boulder: University of Colorado Psychology Dept*. 1977.
17. 厚生労働省研究班, *都市部における認知症有病率と認知症の生活機能障害への対応.* 2013.
18. 呉田陽一, 伏見貴夫, and 佐久間尚子, *言語能力の加齢変化.* 第 9 回東京都老年学会誌, 2002. **9**: p. 200-205.