LN-001

読解支援システムの統一的評価法

小谷 克則†，吉見 毅彦‡，†，九津見 毅††，佐田 いち子††，井佐原 均†
Katsunori Kotani，Takehiko Yoshimi，Takeshi Kutsumi，Ichiko Sata，Hitoshi Isahara

1．まえがき

本稿は，機械翻訳システムや単語訳振りシステムなどを一括して評価する手法を提案する。これまで各システムを個々に評価する手法（例えば，"BLEU" Papineni et al. 2002）は，数多く提案されてきた。これら既存の評価法の多くがシステムの出力の精度を評価尺度として採用するために，出力形態が異なるシステムを統一的に評価することは困難であると思われる。これらの機械翻訳システムや単語約振りシステムを読解支援システムとみなした場合，統一的な評価が必要であると，筆者らは考える。なぜなら，支援システムを統一的に評価すれば，あるテキストに対して有効な支援システムを特定することができるからである。この情報に基づいて，利用者は，機械翻訳や単語訳振りシステムを使い分けることができる。

そこで，筆者らは出力形態に依存しない評価法として，大黒（1993）や富士他（2002）が提案する評価法を採用し，異なる読解支援システムを対象とした評価実験を行った。先行研究（大黒 1993；富士他 2002）が提案する評価手法では，システムの出力結果を実際にユーザーが読解し，そのテキストに対する理解度テストの結果を基にシステムの有効性を評価する。大黒（1993）が対象としたのは単語訳振りシステムであり，富士他（2002）が対象としたのは機械翻訳システムであった。筆者らは有効性に基づく評価手法であれば，単語訳振りシステムと機械翻訳システムを一括して評価できると考え，評価実験を行った。

有効性に基づく機械翻訳システムや単語約振りシステムの評価の尺度として，理解度テストだけでなく，読解速度も用いることを提案する。一般的にテキスト読解の効率を示す要因として，テキスト理解度と読解速度が挙げられる（Alderson 2000）。したがって，理解度テストだけでなく読解速度を尺度として加えることにより読解効率の観点から評価することができる。さらに，後述するように読解速度を評価尺度として採用することにより詳細な評価が可能になると思われる。

本稿の構成は，以下の通りである。まず，次節において有効性に基づく評価法を提案する先行研究を紹介する。3節では，本稿が提案する読解時間・速度に基づく有効性の評価手法を述べる。4節では，筆者らが行なった読解支援システムの有効性に関する評価実験の概要とその結果を述べる。5節では，本稿のまとめと今後の課題を述べる。

2．先行研究（富士他 2002；大黒 1993）

大黒（1993）は，英日単語訳振りシステムの有効性を理解度テストの結果を基に評価する実験を行った。実験に用いたテキストは，TOEIC（the Test of English for International Communication）の文章読解問題テキストであった。そして，理解度を問う設問もこのテキストに用意されているものを利用した。実験に参加した被験者は約50名程度であった。

大黒（1993）は全被験者を読解能力が均一になるように実験群と統制群に二分割した。そして，統制群には支援の施されていない英語原文テキストを与え，実験群には単語訳振りシステムの出力結果を併記した加工テキストを与えた。各被験者群は，与えられたテキストを読み，理解度テストの設問に答えた。この実験を通じて得られた理解度テストの得点から，読解能力が低い人に対して単語訳振りシステムがより有効に機能することが確認された。一方，読解能力が高い人に対してはその有効性は確認されなかった。

次に，機械翻訳システムを対象に実験を行った富士他（2002）を概観する。富士他（2002）も，大黒（1993）と同様に TOEIC の文章読解問題をテキストとして選定した。実験に参加した被験者数は約200名であった。

富士他（2002）は，統制テキストとして英語原文のみのテキストを用意した。そして，実験テキストとして，機械翻訳システムの出力結果のみの翻訳文テキストと出力結果と原文の両方が併記されるテキストの二種類を用意した。この実験を通じて得られた理解度テストの結果から，英語原文テキストと併記文テキストの間で統計的な有意な差が確認された。さらに，富士他（2002）は，大黒（1993）と同様に支援システムの有効性が読解能力の低い群に顕著に現れることを確認した。

大黒（1993）と富士他（2002）は，それぞれ対象としたシステムは異なるが，両者共に支援システムの有効性が被験者の読解能力に応じて変化することを確認した。このことから筆者らは，本稿の提案手法においても同様の効果が確認できれば，読解速度を評価尺度として用いることを支持する証拠の一つになると考えた。本稿では，実験を通じて，支援付きテキストの読解速度が被験者の英文読解能力に応じて変化しているか否かを調査することにより，読解速度による統一的評価法の有効性を検証する。

3．評価指標としての読解速度・時間

支援システムの有効性を評価する尺度して理解度がこれまで主に用いられてきた。理解度テストが検査対象とするのは，テキストの全体，あるいはその一部における読解である。

このような理解度テストの形態で文単位での読解を検査することは，困難であると筆者らは考える。その理由として，文単位で設問を解きながら読解を進めることは，被験者にとってかなり特異な環境であるため，評価の信頼性が低下すると考えられるためである。しかし，読解速度であれば，被験者に大きな負担を与えることなく計測することが可能である。このように読解速度を評価尺度とすると文単位といった局所的に読解の有効性を確認することが可能となる。このような局所的な評価が可能となれば，読解に影響を与える様々な言語特徴（単語，文構造など）を特定することが可能となる。そして，どのような言語特徴が支

援システムにとって問題であるかを特定することも可能になると筆者らは考える。また，理解度テストを用いる場合，評価用テキストに理解度を問う設問が用意されている必要がある。そのため，実験で用いることができるテキストが設問付きのテキストに限定されたり，選定したテキストに設問を作成する必要が生じたりする。一方，読解速度を評価尺度する場合，基本的にどのようなテキストでも対応可能である。したがって，評価対象となるテキスト選定の自由度が高くなる。これらの理由により，筆者らは読解速度を評価尺度として加えることにした。

次に，読解速度を評価尺度として採用することの妥当性を検証するために，筆者らが行なった予備実験の結果を報告する。読解速度を尺度として用いるためには，速度がある程度，テキストの読みやすさを反映している必要があると筆者らは考えた。そこで，筆者らは読解速度がどの程度，テキストの読みやすさと相関があるのかを実験を通じて確認した。

実験で用いたテキストは，本実験で使うテキストと同様の TOEIC の文章読解問題テキストである。このテキストに含まれる文毎の読解速度を計測し，各文の読みやすさを示すスコアと比較した。読みやすさスコアの算出には，いわゆるリーダビリティーの算出式（Flesch 1948）を用いた。実験を通じて，リーダビリティースコアと読解速度の比較により，ある程度の相関（$r=0.7$, $p<0.01$）を確認した。この相関関係を基に，筆者らは読解速度が読みやすさを反映すると想定した。今回，筆者らが用いたリーダビリティの算出式は，英語母語話者を対象に開発されたものである。このような母語話者対象の算出式を英語を第二言語とする読解者に援用することの問題点が指摘されている(新井 2003)。この点に関しては，今後の課題とする。

4. 読解速度・時間に基づく評価実験

4. 1 実験方法

本実験で検証する仮説は次の二つである。

仮説１：支援効果は，英語読解能力低群，中群，高群の順で小さくなる。

仮説２：読解速度は，英語原文テキスト，支援付きテキスト，人間による翻訳テキストの順で速くなる。

分析対象とした支援システムは，（1）チャンキング情報提示システム，（2）単語訳振りシステム，（3）機械翻訳システムであった。実験で用いた読解テキストは，先行研究と同様，TOEIC の文章読解問題テキストから抜粋した 84 テキストであった。また，実験に参加した被験者は 102 名であった。各被験者の英語運用能力を確認するために，被験者全員に TOEIC 試験のスコア票の提出を求めた。被験者の TOEIC スコア分布は表１の通りである。

TOEIC スコア	人数（人）
400-595	36
600-795	36
800-995	30

表１：TOEIC 得点分布

84 個の実験テキストについて，それぞれ支援システムの出力結果を反映させた実験テキスト 4 種類を作成し，合計 336 テキストを用意した。支援タイプの一つが，連語・句などのチャンク情報が示されたチャンキングテキスト（CHU）である。次に，単語訳振りシステムの結果を英語原文テキストに加えた訳振りテキスト（RUB）である。そして，機械翻訳を用いたテキストは，原文と併記したテキスト（Mt-EJ）と翻訳文のみのテキスト（Mt-J）の二種類を用意した。

さらに，実験統制テキストとして，英語原文のみのテキスト（ENG）と人間による翻訳人間による翻訳テキスト（JPN）を用意した。これらの実験・統制テキスト 6 種類を各被験者にランダムに割り振った。英語原文 84 テキストの平均単語数と文数は以下の表の通りであった。

	単語数	文数
平均	89.6	5.9
最小	27	1
最大	266	24
標準偏差	44.3	3.4

表２：テキスト情報

読解時間を計測には，図１に示す読解過程検定システム(吉見他 2005)を用いた。PC 画面上の数字が書かれた文表示アイコン上に，被験者がマウスを用いてカーソルを移動させると，システムは指定された文を表示する。さらに，このアイコンにマーキングを施すことも可能であるため，理解度テストの際，適切な答えと対応したアイコンをマーキングすることにした。

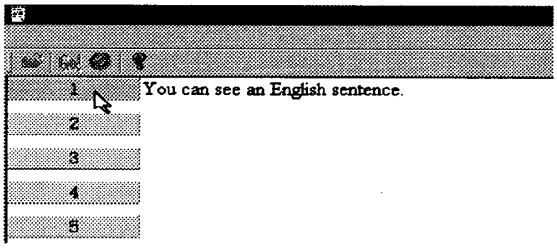

図１：読解過程検定システム

この検定システムを用いて収集された読解時間から速度を算出する際，支援の有無に関わらず，原文に含まれる単語数を基に速度を算出した。本実験は，英語・日本語の混合テキストにおける読解速度の計測が目的ではなく，支援システムを行なうことによりどの程度読みやすくなったかを検出することが目的である。しかし，読解速度の算出時に，実際に表示される単語量，例えば，機械翻訳付きテキストの場合，英語原文に含まれる単語数に加え，翻訳文に含まれる日本語単語数も対象として速度を算出するため支援付きテキストの速度が速くならざるをえない。この影響を排除するために，本実験で速度を算出する際，英語原文に含まれる単語数を用いた。

さらに，本稿では実験で得られたデータの中から特に新聞記事や雑誌記事などだけを分析対象とする。TOEIC テキストには記事以外に広告やメモなど様々なテキストタイプ

が含まれる。筆者らは，これらのテキストタイプの中から支援システムの有効性を確認しやすいと思われるテキストタイプとして記事テキストを選んだ。なぜなら，テキストに含まれる文の数や一文に含まれる単語数が他のテキストに比較して大きいため，読解自体が困難であると思われる。実際，全 84 種類のテキストの平均文数と一文当りの平均単語数を比較すると記事タイプの方が大きいことがわかる。また，記事タイプのテキストは，その他のメモ・告知・宣伝タイプのテキストと比較して，いわゆるリーディング・ストラテジーといいたトップダウンの認知過程よりも言語情報を中心としたボトムアップの過程が占める割合が高いと思われる。本実験の目的が言語情報による支援システムの有効性を評価するため，このような限定を行なった。

	単語数	文数
テキスト全体	89.6	5.9
記事タイプテキスト	142.9	9.6

表３：テキスト全体と記事タイプテキストの比較

4. 2 実験結果と考察

まず，テキスト全体に対して各支援システムの比較評価を行なう。対象となる記事テキストは 13 テキストである。各テキストの平均読解速度を統制テキストである人間による翻訳人間による翻訳テキストと英語原文テキストの速度と比較する。

読解速度の測定結果を表４に示す。表４からわかるように，最も速かったテキストは人間による翻訳テキスト(JPN)であった。最も遅かったテキストは，英語原文テキスト(ENG)ではなく単語訳振りテキスト(RUB)であった。前者は仮説 2 の通りであったが，後者は仮説 2 と異なる結果となった。後者が予測と異なった理由の一つとして，筆者らは単語約振りテキストに含まれるシステムのエラーにより読解が妨げられたことが一つの要因ではないかと考える。同様のことは英語テキストよりも読解速度が遅かったチャンキングテキストにもいえると考える。一方，機械翻訳テキストもいくつかのエラーが含まれるが，エラーによる妨げ以上の支援効果が発揮されたため英語原文テキストの速度より速くなったと考える。

	n	Mean(WPM)	SD	SE	95% CI of Mean
ENG	152	75.16	31.89	2.58	70.0 to 80.2
CHU	147	74.19	36.45	3.00	68.2 to 80.1
RUB	151	65.56	28.00	2.27	61.0 to 70.0
MT-J	146	102.55	56.91	4.71	93.2 to 111.8
MT-EJ	159	70.25	31.73	2.51	65.2 to 75.2
JPN	140	163.13	80.70	6.82	149.6 to 176.6

表４：テキスト読解速度比較

上記の被験者全体の分析結果をさらに被験者の TOEIC 得点群ごとにみていく。被験者を TOEIC スコアに応じて，(1)400－595，(2)600－795，(3)800－995 に分類し，この得点群毎に読解速度を比較した。結果を表５に示す。表５からわかるように，最低速度と最高速度を示したテキストは，被験者を得点群でわけない場合と同様であった。また，英語原文テキスト（ENG）と支援付きテキストを比較する。まず，単語訳振りテキスト(RUB)の場合，全ての能力群において支援効果はみられないが，速度の下降の程度は中群が最大である。したがって，仮説１は支持されなかった。機械翻訳テキスト(MT-J)の場合，英語テキスト(ENG)と比較して，全ての能力群において速度の上昇が確認でき，上昇の割合は低群が最も高く高群が最も低い。したがって，仮説１は支持された。もう一種の機械翻訳テキスト(MT-EJ)の場合，中群と高群は速度が下降しているのに対し，低群は速度が上昇している。したがって，仮説１は支持されなかった。但し，被験者を低群と中・高群に二分すると仮説１は支持される。

	400-595	600-795	800-995
ENG	62.414	73.198	89.197
CHU	63.195	63.389	98.103
RUB	58.383	59.953	80.326
MT-J	109.569	98.613	100.115
MT-EJ	71.447	60.838	80.655
JPN	172.159	152.053	170.922

表５：得点群によるテキスト読解速度比較

この得点群にもとづく平均速度を分散分析により，統制テキストの速度と有意に異なる支援方法を割り出した(表６)。表６において，有意差($p<0.0001$)が確認できたテキストには「√」を付し，確認できなかったテキストには「*」を付した。

	400-595		600-795		800-995	
	ENG	JPN	ENG	JPN	ENG	JPN
CHU	*	√	*	√	*	√
RUB	*	√	*	√	*	√
MT-J	√	√	√	√	*	√
MT-EJ	*	√	*	√	*	√

表６：統制テキストとの読解速度比較

また，各得点群における理解度テストの結果を表７に示す。低得点者群では英語原文テキストの正解率が一番低く，人間による翻訳テキストの正解率が一番高いことがわかる。正解率(表７)と速度(表５・６)を併せて考慮すると，低得点群では統計的に有意差として機械翻訳文のみテキストの有効性が確認できる。また，中得点群では正解率としては機械翻訳文のみテキストが最低であるが，速度としては，人間による翻訳テキストの速度に次ぐ速さである。特に単語訳振りテキストと比較した場合，正解率こそ 0.05 ポイントの差があるにせよ，速度の観点からするとおよそ 40WPM 速くなっていることがわかる。これに対し，高得点群では機械翻訳文のみテキストの正解率は，他のテキストと比べかなり低いのに対し，読解速度はチャンキングとほぼ同程度であることが確認できる。また，どの支援方法であってもその速度において統計的に有意差を確認できるテキストはなかった。

	S1	S2	S3	S4	S5	S6	S7	S8
ENG	78.2 (5)	91.3 (4)	68.9 (5)	78.8 (5)	63.4 (4)	86.7 (2)	76.0 (4)	52.9 (5)
CHU	88.2 (3)	86.7 (5)	83.3 (4)	81.8 (3)	74.1 (3)	78.2 (3)	83.0 (3)	62.7 (3)
RUB	69.2 (6)	60.1 (6)	52.1 (6)	72.9 (6)	57.3 (5)	50.7 (6)	55.1 (6)	57.4 (4)
MT-J	180 (2)	132.9 (2)	98.3 (2)	81.8 (3)	117.8 (2)	76.5 (4)	106.9 (2)	75 (2)
MT-EJ	82.5 (4)	93.7 (3)	95.2 (3)	99.0 (2)	56.8 (6)	74.2 (5)	74.4 (5)	42.1 (6)
JPN	183.6 (1)	259.2 (1)	222.2 (1)	300 (1)	253.8 (1)	218.1 (1)	192.8 (1)	145.9 (1)

表８：読解速度比較（文単位）

ENG	If no one within the firm is qualified, look outside the organization.
CHU	If/ no one/ within the firm/ is/ qualified,/ look/ outside the organization.
RUB	If no one within the firm is qualified, look outside the organization.
MT-J	～の内に 会社 ～に資格を与える ～の外で 組成 会社の中のだれも資格を与えられないならば，組織の外で見なさい。
MT-EJ	If no one within the firm is qualified, look outside the organization. 会社の中のだれも資格を与えられないならば、組織の外で見なさい。
JPN	もし，資格にかなう者が社内で見当たらない場合，社外を当たらなければならない。

表９：例文（S6）

	400-595	600-795	800-995
ENG	0.682	0.891	0.927
CHU	0.736	0.848	0.922
RUB	0.744	0.825	0.921
MT-J	0.766	0.820	0.838
MT-EJ	0.867	0.929	0.906
JPN	0.874	0.955	0.937

表７：テキスト理解度（設問正解率）

次に 13 個のテキストの中から読解速度を比較して英語原文テキストと人間による翻訳テキストに有意差が確認できた７つのテキストの中から１テキストを取り出して，文単位で読解速度の比較分析を行なう。分析対象のテキストは８文からなるものであった。各テキストタイプにおけるそれぞれの文の速度は以下の表８の通りである。表８においてカッコ内の数字は，各文における支援方法間の順位を表す。統制テキストである人間による翻訳テキストの速度は，全ての文において１位であった。次に，英語原文テキストの場合，常に最下位ではなかった。S6 における英語原文テキストの速度はどの支援付きテキストよりも速いという結果が得られた。速度が読みやすさを反映するという想定が正しければ，何らかの理由により支援付きテキストの読みやすさが低下したと考えられる。支援システムの誤りにより読みやすさが低下したと筆者らは考えた。そこで，実際に支援付きテキストを比較し，その精度を主観的に確認することにした。

例文(表９)にあるように，S6 には"no one"のように日本語の単語へ直訳が困難と思われる要素がある。特にこの文中では，人間による翻訳文が示すように単に「誰も～をしない」という状態を示すだけではなく，「誰も～しないことがわかる」という意味内容が適切である。さらに，If-節中の"is qualified"は，BE 動詞と一般動詞の過去分詞から成り，その意味が受動態か完了形かが曖昧である。

実際に支援システムの出力をみると機械翻訳システムは，No-one の問題と Be-qualified の問題の双方において誤りであると考えられる。また，単語訳振りシステムも機械翻訳システムと同様に双方の問題において誤りを含む。被験者はこれらの誤りを含むテキストから正しい解釈を復元する必要がある。そしてこの復元に要する時間が計上されることにより読解速度の低下を招くと筆者らは考える。なぜなら，意味的な誤りを含まないチャンキングテキストが機械翻訳システムと単語訳振りシステムのいずれよりも速度が速いからである。

５．まとめと今後の課題

本稿は，機械翻訳システムや単語訳振りシステムなどの異なる読解支援システムを一括して評価する手法を提案し，その評価実験を行った。

仮説１については，被験者を二群に分割した場合，支持されることがわかった。また，仮説２に関しては支援付きテキストとを人間による翻訳テキストと比較した場合は支持されるが，英語テキストと比較した場合は支持されない。仮説１がほぼ支持されることから，読解速度による統一的評価法が有効であることが確認された。

評価実験において，テキスト単位での有効性を評価した場合，理解度テストと同等の評価が読解速度を尺度した場合にも得られることを確認した。また，理解度テストによる評価が困難とする文単位での評価もおこなった。

参考文献

Alderson, J. C.: Assessing Reading. Cambridge University Press: Cambridge (2000)

Flesch, R.: A New Readability Yardstick. Journal of Applied Psychology 32 (1948) 221-233

Fuji, M., N. Hatanaka, E. Ito, S. Kamei, H. Kumai, T. Sukehiro, T. Yoshimi, & H. Isahara Evaluation Method for Determining Groups of Users Who Find MT "Useful." Proceedings of the MT Summit VIII (2001)

Ohguro, Y Evaluating the Validity of Printing Japanese Words alongside English Text. Technical Report on Information Processing Society of Japan, 93-NL-79 (1993) 127-134

Papineni, K., S. Roukos, T. Ward, & W.-J. Zhu BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (2002) 311-318.

新井七菜子.:日本人英語学習者向けのリーダビリティ式の提案と評価に関する研究.第 42 回大学英語教育学会全国大会要綱.(2003) 95-96.

吉見毅彦，小谷克則，九津見毅，佐田いち子，井佐原均.:英語学習の読解能力推定のための読解時間測定法.教育システム情報学会学会誌.Vol.22. 24-29.