TOSHIBA
Leading Innovation

「空気の読める機械翻訳」の評価方法

（株）東芝
研究開発センター　知識メディアラボラトリー
鈴木　博和
hirokaz.suzuki@toshiba.co.jp

東芝グループは、持続可能な
地球の未来に貢献します。

ポイント

- **人手評価で出たスコアを使えば十分？**
- **明確なゴール設定と現在の到達度が分かるようにしたい。**

- **スコアを使えば「相対評価」は可能**
  - （例）従来は**60点**の翻訳品質、改良により**80点**の翻訳品質を実現
  - （例）他社は**60点**の翻訳品質、当社は**80点**の翻訳品質を実現
- **やりたい評価は「絶対評価（到達度評価）」**
  - **満足する翻訳品質を得るためには何点必要なのか？**
  - **現在の翻訳品質はどのようなレベルなのか？**
    - 小学生レベル、中学生レベル、高校生レベル、大学生レベル、一般人レベル、etc.
    - カタコトレベル、非ネイティブレベル、ネイティブレベル、etc.
    - 使えないレベル⇔人の翻訳と区別がつかないレベル

「空気の読める機械翻訳」を作りたい

「空気が読める」とは

- **「ナイフじゃこのケーブルは切れません」**
  - シーン１

    「この細いケーブルは非常に丈夫です。」

    「どのくらい？」

    「ナイフじゃこのケーブルは切れません」

    - *You cannot cut this cable with a knife.*

  - シーン２

    「ナイフじゃこのケーブルは切れません」

    「なぜ？」

    「切れが鈍くなっているからです」

    - *You cannot cut this cable with the knife.*

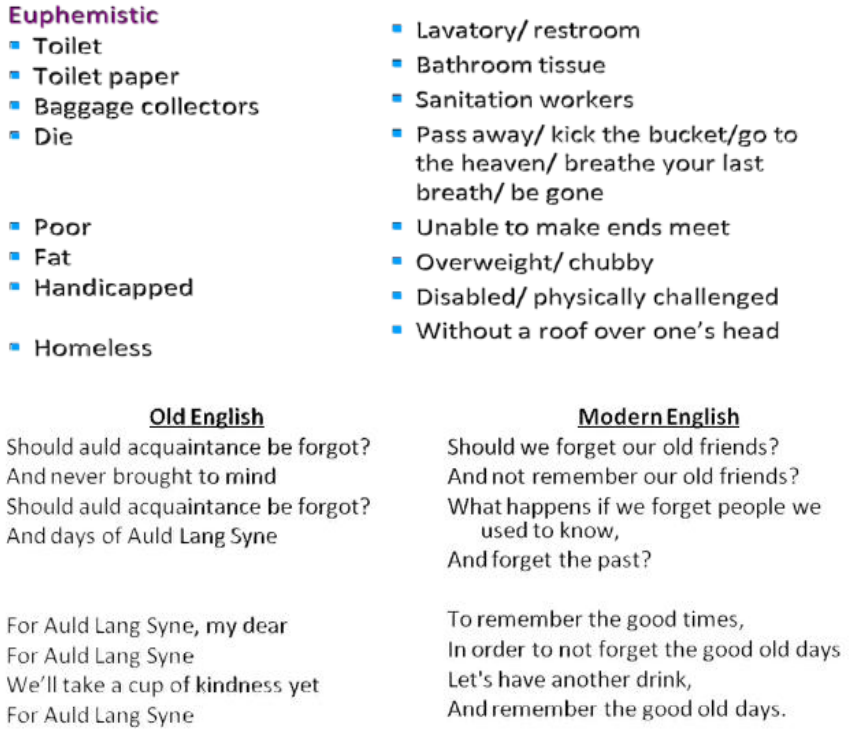
「空気の読める」機械翻訳の解釈

- **工学的**には3種類のadaptation
  - **Domain Adaptation**
    - 対象のドメインに合わせて
  - **Context Adaptation**
    - 文脈、状況、社会的常識などに合わせて
  - **Intention Adaptation**
    - 発話者の意図に合わせ

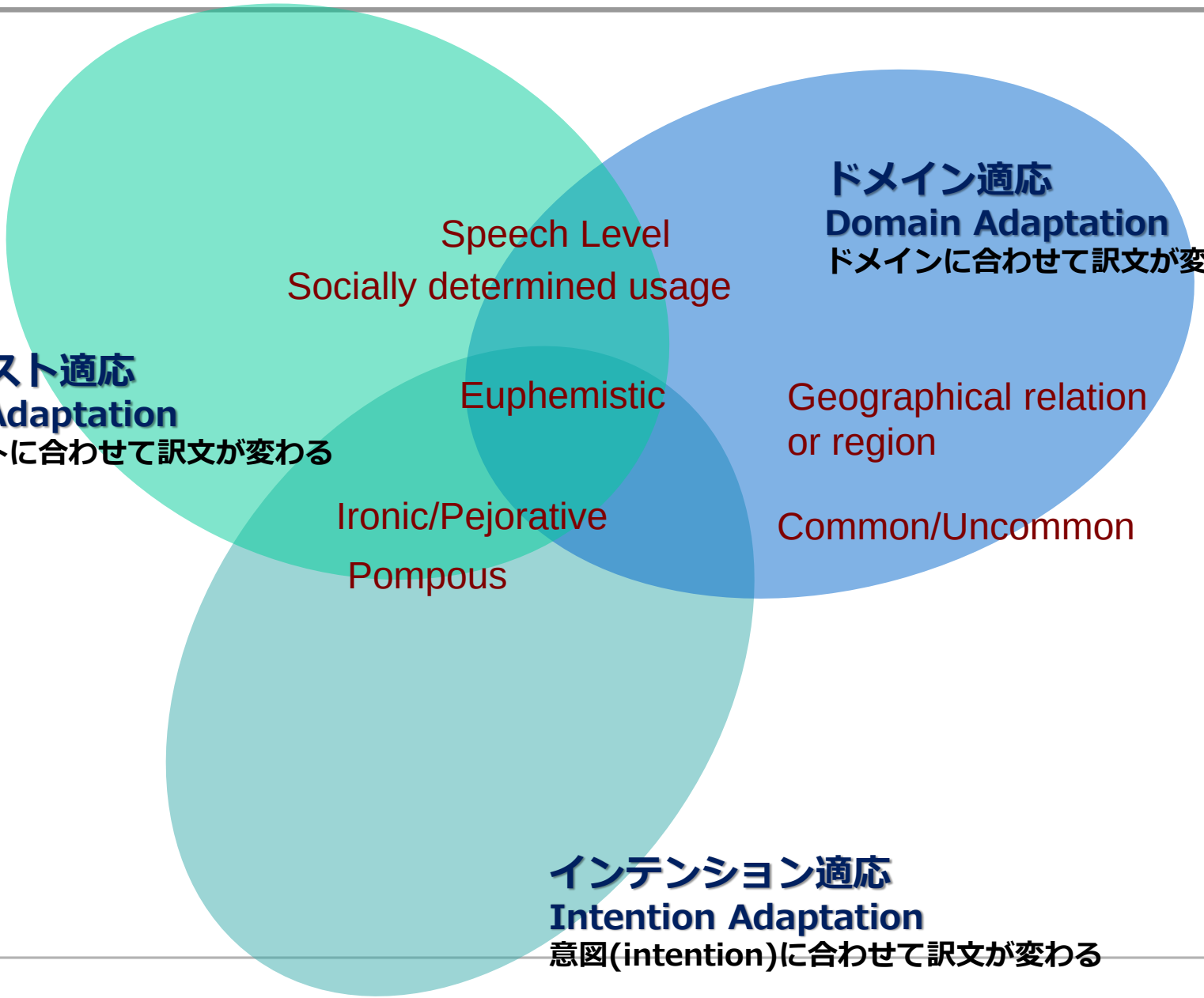
- **翻訳学的には**
  - 「空気の読める」＝**Equivalent Effect** (Nida, 1964)

Koller(1997)のConnotative Equivalence

- **Speech Level**
  - elevated/poetic/normal/colloquial/slang/vulgar
- **Socially determined usage**
  - student/military/aristocratic/…
- **Geographical relation or region**
  - American English/Australian English/…
- **Pompous**
  - “I told you a million times!”
- **Euphemistic**
- **Common/Uncommon**
- **Ironic/Pejorative**
  - ”Lovely day for a picnic!”



AdaptationとEquivalence



問題

- このような「空気の読み具合」を評価する枠組みがない。
- どのような翻訳になればゴールなのかもわからない。

- **翻訳のプロである実務翻訳業界にヒントはあるだろうか？**

実務翻訳での評価は？

実務翻訳での品質評価

- **実務翻訳での品質評価**
  - **Society of Automotive Engineers J2450**
    - 自動車サービスに関する文書の翻訳品質の人手評価metric
    - 人手評価をシステマティックに行い、スコアを算出するので現実性・実用性が高い

J2450　スコアシート

SAE J2450
Translation Metric Score Sheet

Error Type	Num * Serious	Num * Minor	Category Weighted Score
Wrong Term Score WT	_____ * 5 _____ +	_____ * 2 _____ =	_____
Syntactic Error Score SE	_____ * 4 _____ +	_____ * 2 _____ =	_____
Omission Score OM	_____ * 4 _____ +	_____ * 2 _____ =	_____
Word Structure/Agreement Score SA	_____ * 4 _____ +	_____ * 2 _____ =	_____
Misspelling Score SP	_____ * 3 _____ +	_____ * 1 _____ =	_____
Punctuation Error Score PE	_____ * 2 _____ +	_____ * 1 _____ =	_____
Miscellaneous Error Score ME	_____ * 3 _____ +	_____ * 1 _____ =	_____

Document Score: (sum of weighted scores ÷ number of words in source language document)

_____ + _____ + _____ + _____ + _____ + _____ + _____ = _____ Sum of Weighted Scores

_____ Sum of Weighted Scores ÷ _____ Number of Words in Source Text =

_____ ***Overall Document Weighted Score***

（採点例）

Example Category Score in Syntactic Error category with major weight 4 and minor weight 2, assuming 3 major syntactic errors and 4 minor syntactic errors:

3 * 4 + 4 * 2 = 20

Example Document Score: (in a document with 330 source text words)

14 + 20 + 10 + 8 + 9 + 1 + 7 = 69 (Sum of Weighted Scores in all 7 categories)

69 (Sum of Weighted Scores) ÷ 330 (Number of Words in Source Text) =

0.209 ***(Overall Document Weighted Score)***

J2450

APPENDIX B

SAE J2450 QUICK REFERENCE

B.1 See Figure B1.

J2450 Quick Reference

1. ***When an error is ambiguous, always choose the earliest primary category.***
2. ***When in doubt, always choose 'serious' over 'minor.'***

A. Wrong Term: (WT) A 'wrong term' is any target language term that
a. violates a client term glossary;
b. is in clear conflict with de facto standard translation(s) of the source language term in the automotive field;
c. is inconsistent with other translations of the source language term in the same document or type of document unless the context for the source language term justifies the use of a different target language term, for example due to ambiguity of the source language term;
d. denotes a concept in the target language that is clearly and significantly different from the concept denoted by the source language term.

Serious weight: 5; Minor weight: 2

B. Syntactic Error: (SE) A syntactic error comprises the following cases:
a. A source term is assigned the wrong part of speech in its target language counterpart.
b. The target text contains an incorrect phrase structure, e.g. a relative clause when a verb phrase is needed.
c. The target language words are correct, but in the wrong linear order according to the syntactic rules of the target language.

Serious weight: 4; Minor weight: 2

C. Omission: (OM) An error of omission has occurred if:
a. a continuous block of text in the source language has no counterpart in the target language text and, as a result, the semantics of the source text is absent in the translation;
b. a graphic which contains source language text has been deleted from the target language deliverable.

Serious weight: 4; Minor weight: 2

D. Word Structure or Agreement Error: (SA)
a. An error of **incorrect word structure** has occurred if an otherwise correct target language word (or term) is expressed in an incorrect morphological form, e.g. case, gender, number, tense, prefix, suffix, infix, or any other inflection.
b. An error of **agreement** has occurred when two or more target language words disagree in any form of inflection as would be required by the grammatical rules of that language.

Serious weight: 4; Minor weight: 2

E. Misspelling: (SP) A misspelling has occurred if a target language term:
a. violates the spelling as stated in a client glossary,
b. violates the accepted norms for spelling in the target language,
c. is written in an incorrect or inappropriate writing system for the target language.

Serious weight: 3; Minor weight: 1

F. Punctuation Error: (PE) The target language text contains an error according to the punctuation rules for that language.

Serious weight: 2; Minor weight: 1

G. Miscellaneous Error: (ME) Any linguistic error related to the target language text which is not clearly attributable to the other categories listed above should be classified as a miscellaneous error.

Serious weight: 3; Minor weight: 1

FIGURE B1—SAE J2450 QUICK REFERENCE

実務翻訳では

- 実務翻訳では、見た目(ミススペルなど)も含めてシステマティックに評価する枠組みがある。
- 実務翻訳の場合、「**翻訳が正しく行われているか？**」は欧米では重視される。日本の場合「見た目」重視。
- **翻訳の品質に関する評価はない**

- どうやら実務翻訳の世界にも、翻訳の品質を評価する枠組みはないようだ。

人間翻訳・機械翻訳の両方に使える評価方法の構築は可能か？

現状の人手評価の問題点を新たな観点から

人手評価手法の問題点

- 「空気の読める」機械翻訳を評価するためには**【絶対評価（到達度評価）】**が必要
- 現在の人手評価手法は**【相対評価】**となっている。

- 例えば
  - Adequacy・Fluencyの５段階評価
    - 「Adequacy=5, Fluency＝５」≠「人間が訳したような翻訳」
      - 「Adequacy=4/Fluency＝４の文よりもよい」
      - 評価者・文脈・状況によって、評価が変わる
        - 評価が厳しいグループ vs 評価が優しいグループ

相対評価と絶対評価

- ただ訳文のスコアリングをするだけなら「相対評価」（点差で優劣は分かる）。
- **高品質な訳文とは何か？**

→高品質な訳文だと何を以て評価するか？

→何点が高品質な訳文か？

→何点を取れば合格なのかを考え、それへの到達度を評価する。

到達度によって、**【目標に対し現在どのレベルにあるか】**を評価できる。

人間翻訳に対し、機械翻訳がどのレベルにあるのかも分かるはず！

新しい人手評価手法の提案

手法において考慮すべき点

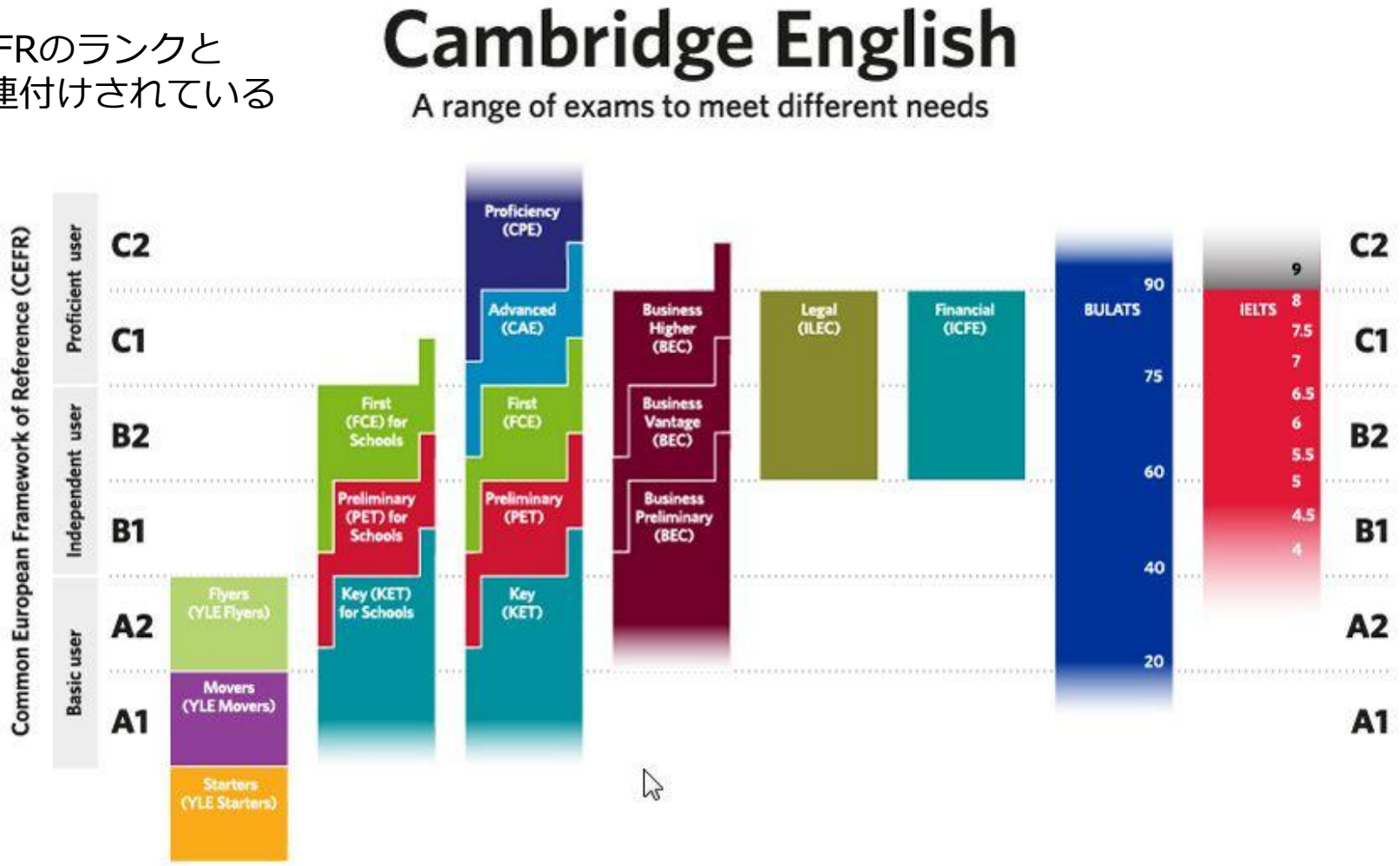
- **主観的な評価を<u>できる限り客観的</u>に扱いたい**
  - 「良い」「悪い」の判断基準
- **評価結果の<u>信頼性</u>、<u>精度</u>、<u>一貫性</u>も検証したい**
  - 評価にバラツキが少ないものがよい
- **「相対評価」ではなく、「<u>絶対評価（到達度評価）</u>」を行いたい**

教育学・心理学・統計学的アプローチも検討。
教育、特に**英語教育の評価方法**にフォーカスしてみる
例えば**Cambridge ESOL**(English for Speakers Of other Language)
Cambridge ESOLのテストは**CEFR**(Common European Framework of Reference)(*)に準拠。
CEFRは**<u>絶対評価（到達度評価）</u>**

(*)日本語版はhttp://wwwsoc.nii.ac.jp/jgg/jggla/library/cef_verzeichnis.html

【参考】Cambridge ESOLとCEFR

CEFRのランクと
関連付けされている



出典:http://www.cambridgeesol.org/about/standards/cefr.html

CEFRとALTE ‘Can Do’ Statementsとの対応

ALTE：Association of Language Testers in Europe

ALTE Can Do Statements: overall general ability			
CEFR LEVELS	Listening/Speaking	Reading	Writing
C2	CAN advise on or talk about complex or sensitive issues, understanding colloquial references and dealing confidently with hostile questions.	CAN understand documents, correspondence and reports, including the finer points of complex texts.	CAN write letters on any subject and full notes of meetings or seminars with good expression and accuracy.
C1	CAN contribute effectively to meetings and seminars within own area of work or keep up a casual conversation with a good degree of fluency, coping with abstract expressions.	CAN read quickly enough to cope with an academic course, to read the media for information or to understand non-standard correspondence.	CAN prepare/draft professional correspondence, take reasonably accurate notes in meetings or write an essay which shows an ability to communicate.
B2	CAN follow or give a talk on a familiar topic or keep up a conversation on a fairly wide range of topics.	CAN scan texts for relevant information, and understand detailed instructions or advice.	CAN make notes while someone is talking or write a letter including non-standard requests.
B1	CAN express opinions on abstract/cultural matters in a limited way or offer advice within a known area, and understand instructions or public announcements.	CAN understand routine information and articles, and the general meaning of non-routine information within a familiar area.	CAN write letters or make notes on familiar or predictable matters.
A2	CAN express simple opinions or requirements in a familiar context.	CAN understand straightforward information within a known area, such as on products and signs and simple textbooks or reports on familiar matters.	CAN complete forms and write short simple letters or postcards related to personal information.
A1	CAN understand basic instructions or take part in a basic factual conversation on a predictable topic.	CAN understand basic notices, instructions or information.	CAN complete basic forms, and write notes including times, dates and places.

CEFRを参考にした人手評価手法の構築

教育学（言語習得）	機械翻訳
言語習得に関する評価ランクを設ける (e.g. CEFR)	**機械翻訳品質**に関する評価ランクを設ける
適切な**criteria**の設定 (e.g. **Can Do Statements**)	適切な**criteria**の設定 (e.g. JEIDAの”機械翻訳システム評価基準”＋**自然さ・流暢さの評価基準**＋**文脈・意図理解に関する評価基準**＋etc.)
問題・得点配分の策定	**品質評価項目・スコア**の策定
適切な**standard**の設定 (何点のスコアがどのランクに対応するか決定)	
Validation （設定したstandardがどれくらい正確に分類できるか確認）	

Step1 – Criteria Setting

- **Criteriaの設定**
  - ランクを決める
    - CFERのCriteriaグリッドを参考にする（A1/A2/B1/B2/C1/C2の6段階）
  - TQをどのように定義するか？
    - **そもそも「翻訳」とは何なのか**を考慮しなければならない。
    - 翻訳学での解釈を工学的に利用できないか？

MTのためのCriteria Gridを作成

- **Equivalent Effectを念頭において、Criteriaを考えてみた。**

（注）

- **言語学・翻訳学者の意見は反映されていません。**
- **私の独断と偏見で作成しました。**

Criteria Grid of MT Quality（仮）

Level		Qualitative Factors				
		Vocabulary Range & Vocabulary Control	Grammatical Accuracy	Socio-Linguistic Appropriateness	Coherence & Cohesion	Fluency & Flexibility
Proficient User Level	C 2	can convey finer shades of meaning precisely	maintains consistent and highly accurate grammatical control	can express fully the socio-linguistic and socio-cultural implications	can create coherent and cohesive texts making full and appropriate use of a variety of organizational patterns	uses expressions with natural, smooth flow (so smooth that the reader is hardly aware of MT)
	C 1	can translate clearly in an appropriate style on a wide range	consistently maintains a high degree of grammatical accuracy (occasional errors in grammar, collocations and idioms)	can use translation flexibly and effectively for social purposes	can produce clear, smoothly flowing, well-structured test	uses expressions with natural, smooth flow
Independent User Level	B 2	has a sufficient range of language to be able to give clear descriptions	high degree of grammatical control (does not make errors which cause misunderstandings)	can avoid crass errors of formulation	can use a number of cohesive devices to link sentences into coherent text (including minor jumpiness)	uses occasional less appropriate expressions
	B 1	has enough language to get by, with sufficient vocabulary	uses reasonably accurately a repertoire of frequently used patterns	-	can link a series of shorter discrete elements into a connected, linear text	the texts are understandable but occasional unclear expressions
Basic User Level	A 2	uses basic sentence pattern	errors may sometimes cause misunderstandings	-	can link groups of words with simple connectors	texts contain expressions which makes the text hard to understand
	A 1	has a very basic repertoire of words and simple phrases	limited control of a few simple grammatical structures	-	can link words or groups of words with very basic linear connectors (e.g. and/then)	text contain expressions which make the text very hard or impossible to understand
Incomprehensible Level	0	impossible to judge	non-grammatical	impossible to judge	impossible to judge	impossible to understand

Step2 – Item Setting

- **TQを評価するための質問事項を決定する**
  - **質問内容は？**
    - Criteria Gridを考慮した評価項目
  - **何段階評価にするか？**
    - 重要度に応じたスコア配分
  - **質問数をいくつにするか？**
    - 評価コスト
- **ここでも言語学者・翻訳者の意見は重要**
  - 何を重視するか？
    - 何を以って翻訳の良し悪しを判断するか？

Step2 – Item Setting

- **評価項目の検証**
  - **Concordance**
    - 評価者同士で意見が割れていないか？
  - **Consistency**
    - 同一評価者の評価一貫性(intra-rater consistency)
    - 評価者間の評価一貫性(inter-rater consistency)
  - **Reliability**
    - 評価項目が知りたいことをちゃんと反映しているか？

Step2 – Item Setting

- **仮に質問・スコアが決定できたとする**
  - （例）7項目10点満点

		Item						
		Q1	Q2	Q3	Q4	Q5	Q6	Q7
Judge	J1	9	8	9	7	8	10	9
	J2	8	7	7	7	8	8	8
	J3	8	7	8	6	6	8	7
	J4	10	8	10	7	8	8	8
	J5	10	9	9	8	8	9	10
	J6	9	9	10	8	7	7	7
	Avg.	9	8	8.8	7.2	7.5	8.3	8.2

①全体の分散を計算

②この分散を計算

- **検証**
  - **Concordance**
    （Statistics for agreement）
    - **Kendall's W**
    - W=0.4815
  - **Consistency**
    - **Intra-class correlation**(②/①)
    - 0.2962（全体の分散の約70%はJudgeの評価の違いによる）
  - **Reliability**
    - **Cronbach's alpha** (内部一貫性（質問事項が調査目的としている特性を測定できているか）を検証する)
      - Alpha=0.83 (一般に0.7以上が好ましいといわれる)
    - **Speaman-Brown formula**(質問数の変化に伴うreliabilityの変化を予想する)

Step3 – Standard Setting

- **Standardの設定**
  - 何点のscoreでどのランクになるか
  - Standard決定手法
    - **Body of Work method**
    - **Logistic Regression**

この辺がstandard?

Table 6.6: Summary of the Rangefinding Round

Folder	Score	A1	A2	B1	B2	Total
1	13	15	0			15
	15	15	0			15
	16	14	1			15
2	18	13	2			15
	19	11	4			15
	21	9	6			15
3	23	10	5			15
	24	7	8			15
	26	5	10			15
4	27	3	10	2		15
	28	0	12	3		15
	30	1	11	3		15
5	32		9	6		15
	33		11	4		15
	34		8	7		15
6	35		7	8		15
	36		8	7		15
	37		6	8	1	15
7	39		3	12	0	15
	41		1	14	0	15
	42		1	12	2	15
8	43			10	5	15
	45			11	4	15
	46			8	7	15
9	48			4	11	15
	49			1	14	15
	51				15	15
10	52				15	15
	53				15	15
	54				15	15

Table 6.7: Results of the Pinpointing Round (partially)

Score	A2	B1	B2	p	$\ln[p/(1-p)]$
32	10	5		0.333	-0.6931
33	11	4		0.267	-1.0116
34	9	6		0.400	-0.4055
35	7	8		0.533	0.1335
36	8	7		0.467	-0.1335
37	6	9		0.600	0.4055
38	4	10	1	0.733	1.0116

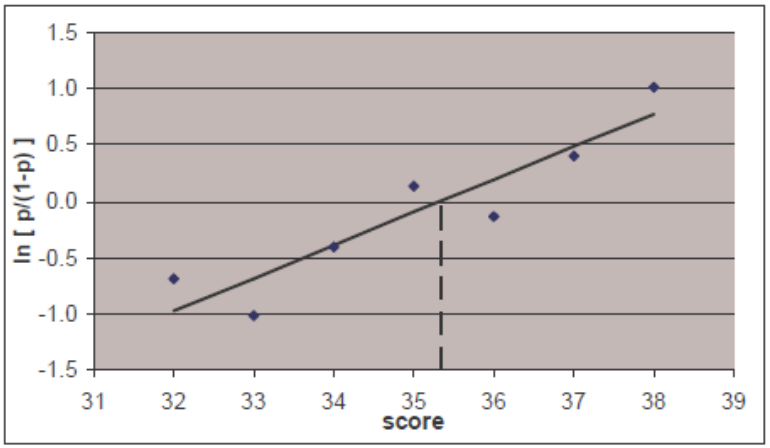


Figure 6.2. Logistic Regression

Step4 – Agreement/Consistency

- **評価者間のAgreement/Consistencyを評価**
- **Agreement**
  - **Cohen's Kappa**(2 評価者間のagreementを検証)
    - Kappa=0.6461
  - **Fleiss' Kappa**（複数評価者間のagreementを検証）
- **Consistency**
  - 順位相関係数
    - **Kendall's tau rank correlation coefficient**
    - **Speaman's rank correlation coefficient**
      - Spearman r=0.6731

Table 7.4: Bivariate Frequency Table using Four Levels

		Judge 2				
		A1	A2	B1	B2	Total
Judge 1	A1	7	2	1	1	11
	A2	1	10	2	1	14
	B1	1	2	12	2	17
	B2	0	1	0	7	8
	Total	9	15	15	11	50

Step5 - Validation

- **設定したStandardのaccuracy/consistencyを検証**
- **評価手法**
  - **Livington & Lewis**の手法を使う
  - ReliabilityをCronbach's alphaで推定
  - ツール（BB-CLASS）を使う
  - Center for Advanced Studies in Measurement and Assessment, Univ. of Iowa

reliability

Table 5: Control Cards and Frequency Distribution for Livingston and Lewis (1995) Example

```
LL 0.9 4    check
"LL data" f 1 2
3  140. 160.      .4   .6
```

設定したstandard

スコア

頻度

121 3	141 5	161 14	181 8
122 5	142 20	162 17	182 3
123 8	143 11	163 17	183 9
124 5	144 14	164 23	184 0
125 3	145 15	165 29	185 7
126 9	146 21	166 19	186 5
127 2	147 13	167 16	187 0
128 2	148 12	168 33	188 2
129 9	149 10	169 12	189 1
130 18	150 18	170 34	190 1
131 10	151 18	171 16	
132 11	152 17	172 21	
133 13	153 8	173 17	
134 12	154 21	174 32	
135 10	155 6	175 0	
136 11	156 33	176 32	
137 16	157 32	177 22	
138 11	158 7	178 14	
139 16	159 17	179 8	
140 15	160 36	180 25	

Step5 - Validation

Table 8: Control Cards and Frequency Distribution for Livingston and Lewis (1995) Example (continued)

```
***ACCURACY RELATIVE TO EXPECTED OBSERVED SCORES GIVEN MODEL***

           x0        x1        x2      marg
  t0   0.14951   0.01090   0.00000   0.16042
  t1   0.06173   0.20016   0.01146   0.27335
  t2   0.00045   0.12324   0.44255   0.56624

marg   0.21169   0.33430   0.45401   1.00000

probability of correct classification = 0.79222
false positive rate = 0.02236;  false negative rate =  0.18542

***CONSISTENCY USING EXPECTED OBSERVED SCORES GIVEN MODEL***

           x0        x1        x2      marg
  x0   0.16625   0.04479   0.00066   0.21169
  x1   0.04479   0.22121   0.06830   0.33430
  x2   0.00066   0.06830   0.38505   0.45401

marg   0.21169   0.33430   0.45401   1.00000

pc = 0.77251;  pchance = 0.36269;  kappa = 0.64304
probability of misclassification = 0.22749

***ACCURACY RELATIVE TO ACTUAL OBSERVED SCORES***

           x0        x1        x2      marg
  t0   0.15114   0.01021   0.00000   0.16135
  t1   0.06240   0.18741   0.01194   0.26174
  t2   0.00046   0.11539   0.46106   0.57690

marg   0.21400   0.31300   0.47300   1.00000

probability of correct classification = 0.79961
false positive rate = 0.02215;  false negative rate =  0.17824

***CONSISTENCY USING EXPECTED (row) VS. ACTUAL (column) OBSERVED SCORES***

           x0        x1        x2      marg
  x0   0.16806   0.04193   0.00068   0.21068
  x1   0.04527   0.20712   0.07116   0.32355
  x2   0.00066   0.06395   0.40116   0.46577

marg   0.21400   0.31300   0.47300   1.00000

pc = 0.77634;  pchance = 0.36667;  kappa = 0.64685
probability of misclassification = 0.22366
```

Absolute agreement

Consistency

Step6 – Test Set

- **評価用セットをどのように準備するか？**
  - ドメイン
  - 文数
  - 文の長さ
  - 含まれている言語現象
- **原文の品質を考慮すると、Readabilityという観点も必要**
  - Memphis大の**Coh-Metrix**

【参考】Coh-Metrix

- **テキストの一貫性や難易度を数値化**
- **数理言語学(computational linguistics)と心理言語学(psycholinguistics)の観点で評価可能**

【参考】Coh-Metrixの出力インデックス

No.	Description	Measure	Full description
1	Title	Title	Title
2	Genre	Genre	Genre
3	Source	Source	Source
4	JobCode	JobCode	JobCode
5	LSASpace	LSASpace	LSASpace
6	Date	Date	Date
7	Adjacent anaphor reference	CREFP1u	Anaphor reference, adjacent, unweighted
8	Anaphor reference	CREFPau	Anaphor reference, all distances, unweighted
9	Adjacent argument overlap	CREFA1u	Argument Overlap, adjacent, unweighted
10	Argument overlap	CREFAau	Argument Overlap, all distances, unweighted
11	Adjacent stem overlap	CREFS1u	Stem Overlap, adjacent, unweighted
12	Stem overlap	CREFSau	Stem Overlap, all distances, unweighted
13	Content word overlap	CREFC1u	Proportion of content words that overlap between adjacent sentences
14	LSA sentence adjacent	LSAassa	LSA, Sentence to Sentence, adjacent, mean
15	LSA sentence all	LSApssa	LSA, sentences, all combinations, mean
16	LSA paragraph	LSAppa	LSA, Paragraph to Paragraph, mean
17	Personal pronouns	DENPRPi	Personal pronoun incidence score
18	Pronoun ratio	DENSPR2	Ratio of pronouns to noun phrases
19	Type-token ratio	TYPTOKc	Type-token ratio for all content words
20	Causal content	CAUSVP	Incidence of causal verbs, links, and particles
21	Causal cohesion	CAUSC	Ratio of causal particles to causal verbs (cp divided by cv+1)
22	Intentional content	INTEi	Incidence of intentional actions, events, and particles.
23	Intentional cohesion	INTEC	Ratio of intentional particles to intentional content

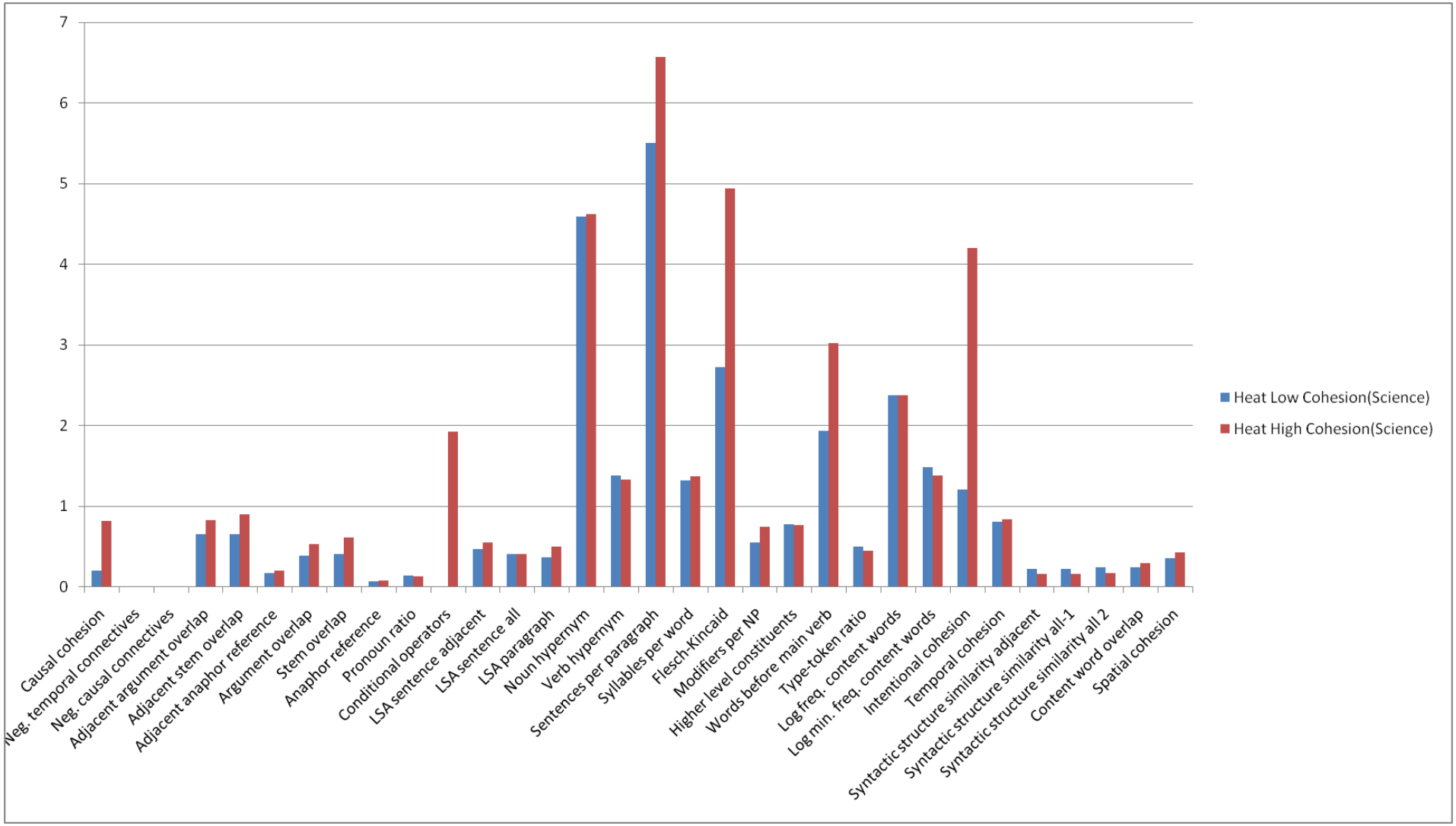
【参考】Coh-Metrixの出力インデックス（続き）

24	Syntactic structure similarity adjacent	STRUTa	Sentence syntax similarity, adjacent
25	Syntactic structure similarity all-1	STRUTt	Sentence syntax similarity, all, across paragraphs
26	Syntactic structure similarity all 2	STRUTp	Sentence syntax similarity, sentence all, within paragraphs
27	Temporal cohesion	TEMPta	Mean of tense and aspect repetition scores
28	Spatial cohesion	SPATC	Mean of location and motion ratio scores.
29	All connectives	CONi	Incidence of all connectives
30	Conditional operators	DENCONDi	Number of conditional expressions, incidence score
31	Pos. additive connectives	CONADpi	Incidence of positive additive connectives
32	Pos. temporal connectives	CONTPpi	Incidence of positive temporal connectives
33	Pos. causal connectives	CONCSpi	Incidence of positive causal connectives
34	Pos. logical connectives	CONLGpi	Incidence of positive logical connectives
35	Neg. additive connectives	CONADni	Incidence of negative additive connectives
36	Neg. temporal connectives	CONTPni	Incidence of negative temporal connectives
37	Neg. causal connectives	CONCSni	Incidence of negative causal connectives
38	Neg.logical connectives	CONLGni	Incidence of negative logical connectives
39	Logic operators	DENLOGi	Logical operator incidence score (and + if + or + cond + neg)
40	Raw freq. content words	FRQCRacw	Celex, raw, mean for content words (0-1,000,000)
41	Log freq. content words	FRQCLacw	Celex, logarithm, mean for content words (0-6)
42	Min. raw freq. content words	FRQCRmcs	Celex, raw, minimum in sentence for content words (0-1,000,000)
43	Log min. freq. content words	FRQCLmcs	Celex, logarithm, minimum in sentence for content words (0-6)
44	Concreteness content words	WORDCacw	Concreteness, mean for content words
45	Min. concreteness content words	WORDCmcs	Concreteness, minimum in sentence for content words
46	Noun hypernym	HYNOUNaw	Mean hypernym values of nouns
47	Verb hypernym	HYVERBaw	Mean hypernym values of verbs
48	Negations	DENNEGi	Number of negations, incidence score
49	NP incidence	DENSNP	Noun Phrase Incidence Score (per thousand words)
50	Modifiers per NP	SYNNP	Mean number of modifiers per noun-phrase

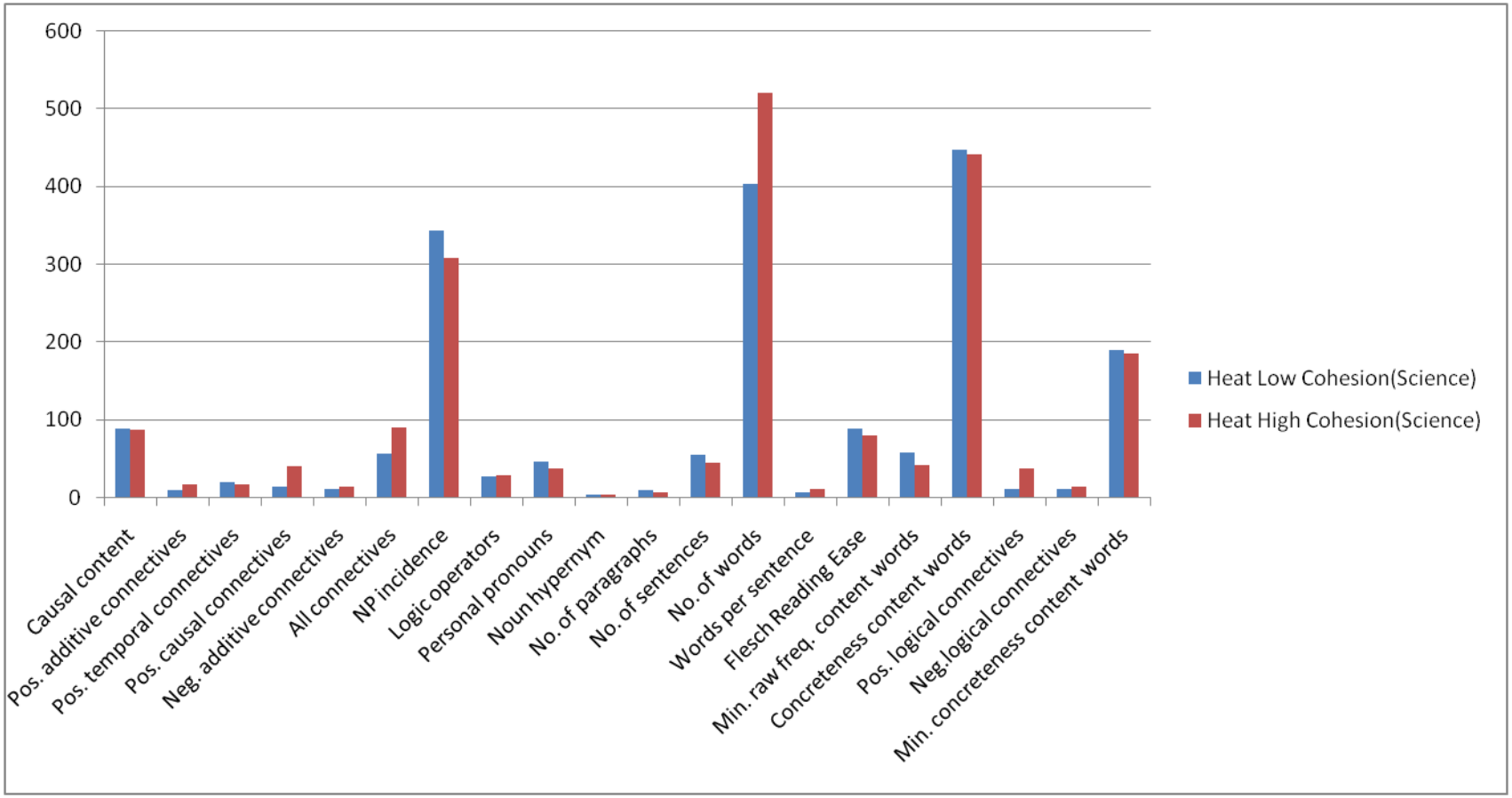
【参考】Coh-Metrixの出力インデックス（続き）

51	Higher level constituents	SYNHw	Mean number of higher level constituents per word
52	Words before main verb	SYNLE	Mean number of words before the main verb of main clause in sentences
53	No. of words	READNW	Number of Words
54	No. of sentences	READNS	Number of Sentences
55	No. of paragraphs	READNP	Number of Paragraphs
56	Syllables per word	READASW	Average Syllables per Word
57	Words per sentence	READASL	Average Words per Sentence
58	Sentences per paragraph	READAPL	Average Sentences per Paragraph
59	Flesch Reading Ease	READFRE	Flesch Ease Score (0-100)
60	Flesch-Kincaid	READFKGL	Flesch-Kincaid Grade Level (0-12)

【参考】Coh-Metrixを実際に使ってみる



【参考】Coh-Metrixを実際に使ってみる（続き）



まとめ

評価手法確立のためにやらなければならないこと

- **人手評価で出たスコアを使えば十分？**
- **なぜ、わざわざランクを設定するのか？**
- **なぜ、わざわざスコアをランクを対応づけるのか？**

- **スコアを使えば「相対評価」は可能**
  - （例）従来は**60点**の翻訳品質、改良により**80点**の翻訳品質を実現
  - （例）他社は**60点**の翻訳品質、当社は**80点**の翻訳品質を実現
- **やりたい評価は「絶対評価（到達度評価）」**
  - **満足する翻訳品質を得るためには何点必要なのか？**
  - 現在の翻訳品質はどのようなレベルなのか？
    - 小学生レベル、中学生レベル、高校生レベル、大学生レベル、一般人レベル、etc.
    - カタコトレベル、非ネイティブレベル、ネイティブレベル、etc.
    - 使えないレベル⇔人の翻訳と区別がつかないレベル

TOSHIBA
Leading Innovation

ご清聴ありがとうございました。

TOSHIBA eco style

Toshiba Group contributes to
the sustainable future of planet Earth.