

Measuring Accuracy

Mathias Schulze & Peter Wood

University of Waterloo & University of Saskatchewan

<http://www.wcgs.ca/~mschulze>

<http://artsandscience.usask.ca/languages/FacultyDetail.php?bioid=1363>

Outline

- **Accuracy** as a dimension of proficiency
- **Accuracy** as gradient feature of constructions
- **Accuracy** and statistical NLP in ICALL – related work
- **Accuracy** indicators

Akkurat	Mono	Number Two	Ex
Replica	Mono	Akkurat	Wa
TABLETTENSCHRIFT			Co
TERMINAL	Paragon	Thermo	L
Simple	Köln-Bonn	Brauer Neue	
Replica	LIQUID CRYSTAL		



CAF - Proficiency in written texts

- **proficiency**(complexity, accuracy, fluency)
- increased **fluency**: longer text for a given task or time period
- increased **accuracy**: more constructions meet readers' expectations
- increased **complexity**: constructions are less predictable due to increased diversity and sophistication

Brief excursion: complexity

$$GTTR = T/\sqrt{W} \quad \text{with} \quad 1/\sqrt{W} \leq GTTR \leq \sqrt{W}$$

$$MWL = L/W \quad \text{with} \quad 1 \leq MWL \leq L$$

$$UBR = U/\sqrt{(W-1)} \quad \text{with} \quad 1/\sqrt{(W-1)} \leq UBR \leq \sqrt{(W-1)}$$

$$MPL = W/P \quad \text{with} \quad 1 \leq MPL \leq W$$

$$z(X) = (X - M(X))/STDEV(X)$$

$$CB = z(z(GTTR) + z(MWL) + z(UBR) + z(MPL)) - STDEV(GTTR, MWL, UBR, MPL)$$

L = number of letters; P = number of period units; T = number of word form types; U = number of bigram types; W = number of word forms; GTTR = Guiraud's type token ratio of word forms; MPL = mean period unit length; MWL = mean word length; UBR = unique bigram ratio; CB = balanced complexity; z(X) = z-score of X (standardized); M(X) = mean of X; STDEV(X) = standard deviation of X

Balanced complexity

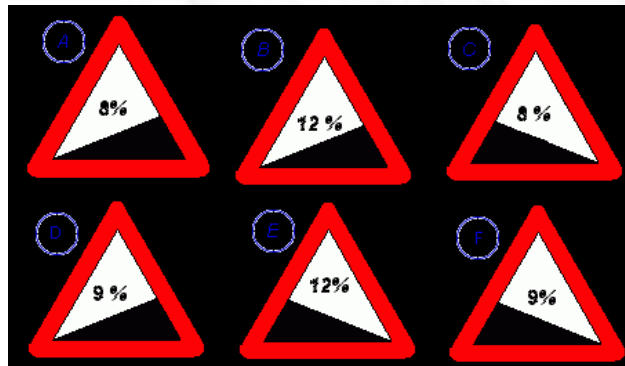


- a robust measure with **concurrent validity**
- 2,072 German texts (186,608 word forms) by 231 elementary/intermediate students from a variety of 43 tasks: **balanced complexity identifies level** (one-way ANOVA (df(2, 2069), $F=355.52$, $p<0.001$); non-parametric Kruskal-Wallis rank sum test ($\chi^2= 481.34$, $p< 0.001$); Tukey Multiple Comparison of Means ($p<0.001$))
- 489 English texts (42,546 word forms) by Dutch school students on two tasks: **balanced complexity differentiates holistic proficiency scores** (one-way ANOVA (df(3, 484), $F= 67.9$, $p<0.001$), non-parametric Kruskal-Wallis rank sum test ($\chi^2= 160.48$, $p< 0.001$); Tukey Multiple Comparison of Means ($p<0.001$ and $p<0.01$))

Gradient (grammatical) accuracy

Grammar as emergent knowledge of conventionalized meaning-form pairings (CG, CxG); lexicon-syntax continuum

Accuracy is not binary, but on a continuum (e.g., degrees of grammaticality)



Aarts (2007): subsecutive (within category: e.g., more or less accurate) and intersective **gradience** (between categories: e.g., both inaccurate and accurate)

Gradient grammatical accuracy can be ranked and is cumulative (Sorace & Keller, 2005)

Gradient grammatical accuracy

Constructions are more likely to be perceived as **accurate** if

- we have seen them before
- they follow a pattern we have seen before

Constructions are more likely to be perceived as **less accurate** if

- they have fewer constituents that are perceived as accurate
- they have constituents that have a very low probability of accuracy



Measuring Accuracy in SLA

In SLA various measures of accuracy have been proposed to quantify task performance and proficiency level



SLA: Accuracy Measures

- Error ratio

$$\phi_t = \frac{e_t \times 100}{w_t}$$

the number of errors divided by the number of words multiplied by 100

- Sentence ratio

$$\phi_t = \frac{\sigma_e}{\sigma_t}$$

the number of erroneous sentences divided by the total number of sentences

Problems



- What is a **word**?
- What is a **sentence**?
- What to do with **multiple instances** of the same error?
- Calculation of these ratios rely on **manual annotation** (training of annotators, resource intensive)

Measuring accuracy in CALL

- Parsing
- “Shallow” parsing
- Use of **constraints** and **mal-rules**
- Focus primarily on locating individual norm violations and providing contingent **feedback** to learners
- Relatively little use of **statistical NLP** in CALL and ICALL



Related work in Statistical NLP

- [Automated Essay Scoring](#)
(Chodorow, Burstein 2004)
- Automatic detection of lexical and syntactic errors using [corpora](#)
- Automatic detection of lexical and syntactic errors using [web queries](#)
(Gamon and collaborators)



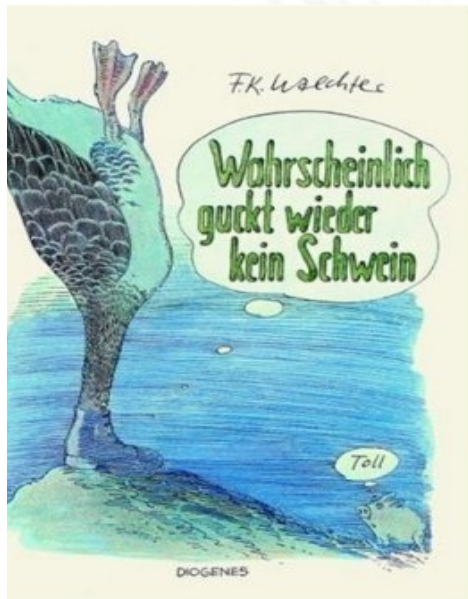


Gradient accuracy indicators



Getting started: some probabilities in a corpus

- For large enough corpora, the number of bigrams (n_{ab}) and trigrams (n_{abc}) is almost identical to the number of tokens in the corpus (n): $n = n_{ab} = n_{abc}$
- If an n-gram is not found in the corpus, we set $n_{n\text{-gram}} = 1$, because we found it in the student text



$$p(abc) = n_{abc} / n$$

$$p(ab) = n_{ab} / n$$

$$p(c | ab) = p(abc) / p(ab) \\ = (n_{abc} / n) / (n_{ab} / n) = n_{abc} / n_{ab}$$

Calculating probable accuracy

- Probable accuracy of a **trigram** (abc):
 - trigram found $\rightarrow \text{acc}(abc) = 1$
 - bigram found $\rightarrow \text{acc}(abc) = p(c | ab) = 1 / n_{ab}$
 - bigram not found $\rightarrow \text{acc}(abc) = p(a,b,c) = n_a * n_b * n_c / n^3$
- Probable accuracy of a **sentence**
 - sentence found $\rightarrow \text{acc}(\text{sent}) = 1$
 - sent not found $\rightarrow \text{acc}(\text{sent}) = \text{rms}(\text{acc}(abc_1) : \text{acc}(abc_s))$

$$= \sqrt{\sum_{i=1}^s \text{acc}(abc_i)^2}$$

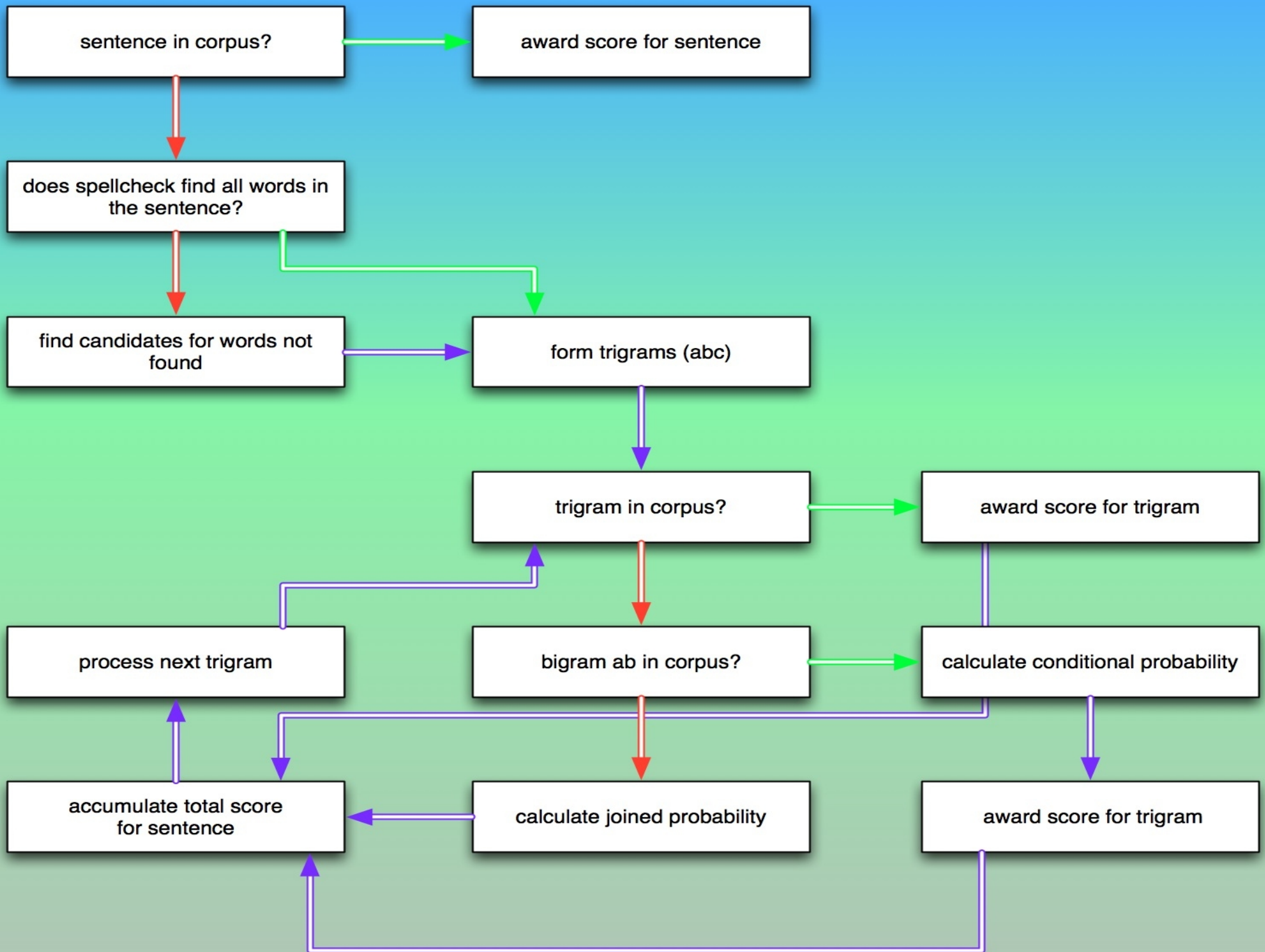


Probable text accuracy

Root Mean Square is the square root of the arithmetic mean of the squared values. Of the different means, it is the one most **tolerant to outliers** and it gets influenced by many smaller deviations (“ganging-up”)

$$\text{acc}(\text{text}) = \text{rms}(\text{acc}(\text{sent}_1) : \text{acc}(\text{sent}_t))$$

$$= \sqrt{\sum_{i=1}^t \text{acc}(\text{sent}_i)^2}$$



Work done

Complexity and **fluency** of written texts can be measured automatically

Medium-sized learner **corpus** compiled and pre-processed

Subset of corpus **hand-annotated** for gradient accuracy of period units

Measures and algorithm proposed

Work to do

Clean and pre-process L1 **corpora** (Google: Web 1T, Wortschatz)

Implement and **test** algorithm

Validate measures

Consider **semantic** accuracy and **task appropriateness**

