Supplementary Material:
Adaptation of Conversations to Communicaton Media

Christian Alis and May T. Lim

National Institute of Physics,

University of the Philippines Diliman

1101 Quezon City, Philippines

MATERIALS AND METHODS

Project Gutenberg (PG)

Author selection

The catalog [1] of Project Gutenberg is available as an XML file that contains records of publications and files linked to each publication. A publication may be available in several file formats with each file size stored in a <`dcterms:extent`> element.

Using Project Gutenberg's catalog released on Oct. 21, 2010, authors were selected based on the amount of their available works in Project Gutenberg, which were estimated by summing the file sizes of plain text, us-ascii-encoded files associated with the author. The first 46 authors were selected by first sorting all authors by decreasing total file size then, starting from the author with the largest file size, an author was selected if he/she appears to have many fictional works with many utterances available. The remaining four authors were selected because it is our impression that their works contain a lot of utterances and were thus investigated during preliminary work.

Parsing

The newline separator of a file was first converted to a single newline character (ASCII code 10, \n). Boiler texts were removed in the beginning and end of each Project Gutenberg file, then quotation mark errors were detected and fixed by first counting the number of quotation marks in a paragraph. If it is even, then there is no quotation mark error but if it is odd, the paragraph is assumed to be the start of a multi-paragraph quote. In the latter case, succeeding paragraphs were processed as follows. If the paragraph begins with a quotation mark, the mark is removed because it implies that the paragraph is part of a multiparagraph quote, otherwise, we leave it as it is. If the number of quotation marks, after possibly removing the beginning quotation mark, is odd then the multiparagraph quote is closed.

Brackets ([]), angle brackets (<>) and text inside them are removed because these are editor's notes or html tags.

Text blocks are considered indented if each line begins with at least two spaces. These blocks typically contain contents of letters or lyrics of songs and is considered as part of an utterance if it occurs inside a multiparagraph quote.

Two successive utterances in the same paragraph are concatenated if the second utterance begins with a lowercase letter to handle utterances split into two by a speaker specifier ("Yes," answered Alice, "that is correct.").

Text between quotation marks were considered as utterances. PG was converted to PGS by splitting each quote in PG over period (.), exclamation point (!) and question mark (?). No spellchecking and normalization of text was performed.

After initial processing, we identified books that contain zero-length utterances. These zero-length utterances are usually due to unclosed quotation marks and use of single quotation marks (') instead of double quotation marks ("). The quotation mark errors in these books were manually fixed then reprocessed until the number of zero-length utterances in the entire PG corpus is negligibly small (0.01 % of the unprocessed data set).

Determination of publication years

A tentative publication year of each book was assigned by searching the US Library of Congress catalog using the query `dc.title all "`*`normalized_title`*`" and dc.creator any `*`surname`* via its Z3950 gateway. The title in the query was normalized by replacing characters that are non-alphanumeric with a space character. From the maximum of 100 results, the earliest year that falls between the author's birthyear and deathyear is selected as the tentative publication year.

The publication year of each book was then manually verified by looking up the copyright page preview of the book in Google Books (`books.google.com`). We used this publication year in the analysis if it is earlier than the tentative publication year from Library of Congress. The publication years of 5.57% of the books were updated this way and the percentage is similar to the estimated publication year error rate of Google Books [2].

Twitter

Data collection intervals

Data was collected using Twitter's spritzer data feed. The first one-week data set was selected randomly while the remaining four were selected due to the notable event happening within the interval as shown in Table I.

TABLE I. One-week data sets

Inclusive dates	Non-zero utterances	Notable event
Sept. 12-18, 2009	5,008,242	
May 8-14, 2010	14,628,664	Philippine elections (May 10, 2010)
Nov. 14-20, 2009	6,782,180	Manny Pacquiao vs Miguel Cotto boxing fight (Nov. 14, 2009)
Dec. 5-11, 2009	7,540,501	Avatar UK release (Dec. 10, 2009)
June 5-11, 2010	18,130,004	World Cup (June 11, 2010)

Parsing

A tweet is considered a reply if it begins with an @ sign. The message of a tweet is obtained by removing all leading usernames using the Python regular expression: `((^| )@\S*)+\S(?=( |$))`. Messages are usually separated from leading usernames by a single space but the given regular expression does not remove this space, thus, typical messages would then be at least of length two.

Subtitles

Parsing

Our complete subtitles dataset consists of all English subtitles in `opensubtitles.org` and was obtained by contacting the website's administrator. Only movie subtitles that are either in srt or sub format were processed. The latest uploaded subtitle is chosen for processing in case a movie has more than one available subtitle. Movies that contain only one utterance implies an encoding error so these were also excluded.

Using the pysubtitles [3] module, the text of each subtitle were extracted then converted into lines of utterances by doing the following, in order, for each line:

1. remove italics and other html markup
2. append line beginning with lowercase into previous line
3. remove hyphens indicating multiple speakers in a frame
4. remove leading whitespaces

This algorithm considers a line as being part of a previous line's utterance if the line begins with a lowercase character. If it begins with a hyphen then the line is a single-line utterance.

Robustness of median median utterance length to author style

Resampling all PG and PGS utterances irrespective of authors into 50 subsets will result to a very sharp median utterance length distribution compared to the original median utterance length distribution (Fig. 1). The resampled subset sizes are not constant and were assigned to correspond with the number of utterances per author in the original datasets. As shown in Fig. 6 in the manuscript, setting each subset sample size equal to some constant large number would only make the resampled median utterance length distribution sharper. The resampled and original median utterance length distributions are not the same suggesting that the utterances have some dependence on each author. However, the median median utterance length of the resampled and original datasets are off by about 1% only. We can therefore use median median utterance length for comparison between data sets.

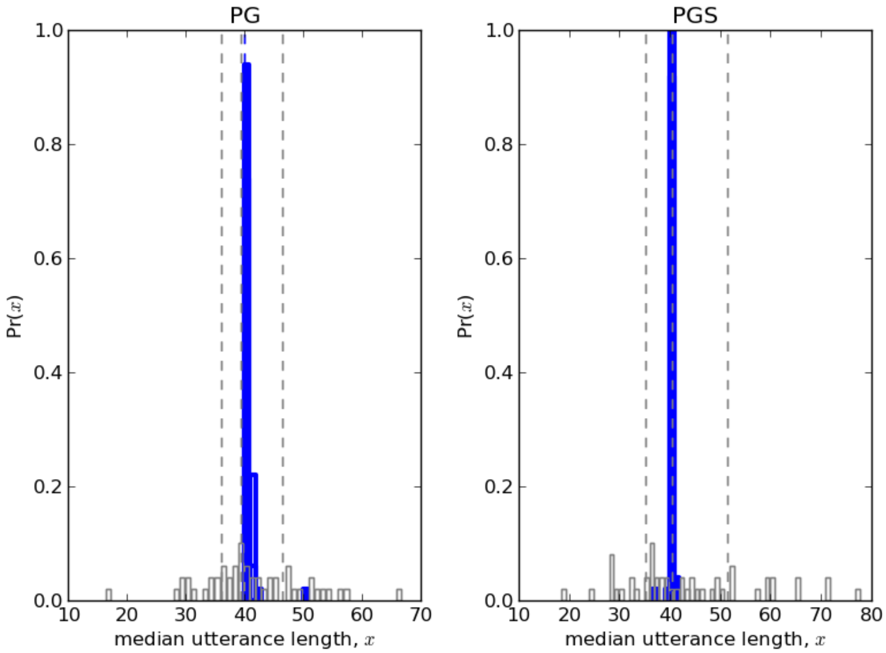

FIG. 1. Median utterance length distribution of original (gray bars) and resampled (blue bars) PG (left) and PGS (right) datasets. Quartiles (gray broken lines) of the original dataset and median (blue broken lines) of resampled dataset are also shown.

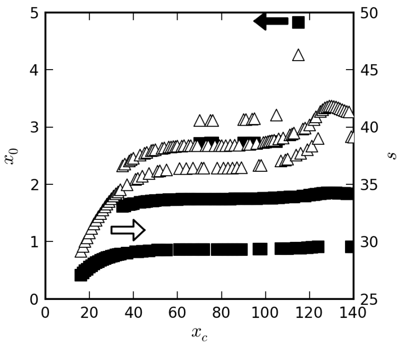

FIG. 2. Best fit location x_0 (filled squares) and scale s (unfilled triangles) parameters when fitting from utterance length $x = 0$ to x_c

RESULTS

Distribution of free parameters

Test results

Sample sizes as in PG

Each data set was resampled to 50 subsets having the same sample size distribution as in PG. The medians of these samples were then compared using Mann-Whitney U test ($n_1 = n_2 = 50$).

TABLE II. Mann-Whitney U test results among datasets

	PG		PGS		SUBS	
	U	p (one-sided)	U	p (one-sided)	U	p (one-sided)
TWITTER	0	3.80×10^{-19} [a]	93	6.11×10^{-17} [a]	0	1.67×10^{-20} [a]
PG			0	1.13×10^{-19} [a]	0	8.29×10^{-21} [a]
PGS					0	4.08×10^{-21} [a]

[a] significant ($\alpha = 0.05$) for both one- and two-sided p-value

Samples sizes equal to N

Each data set was resampled to 50 N-length samples. The medians of these samples were then compared using Mann-Whitney U test ($n_1 = n_2 = 50$).

TABLE III. Mann-Whitney U test results between Twitter and PG

N	U	p (one-sided)
10^2	731	1.74×10^{-4} [a]
10^3	200	2.01×10^{-13} [a]
10^4	0	9.00×10^{-19} [a]
10^5	0	8.09×10^{-21} [a]
10^6	0	5.17×10^{-23} [a]

[a] significant ($\alpha = 0.05$) for both one- and two-sided p-value

TABLE IV. Mann-Whitney U test results between subtitles and Twitter

N	U	p (one-sided)
10^2	440.5	1.18×10^{-8} [a]
10^3	50	3.85×10^{-17} [a]
10^4	0	9.12×10^{-20} [a]
10^5	0	2.70×10^{-21} [a]
10^6	0	5.17×10^{-23} [a]

[a] significant ($\alpha = 0.05$) for both one- and two-sided p-value

TABLE V. Mann-Whitney U test results between subtitles and PG

N	U	p (one-sided)
10^2	0	3.29×10^{-18} [a]
10^3	0	1.72×10^{-18} [a]
10^4	0	4.11×10^{-20} [a]
10^5	0	5.17×10^{-23} [a]
10^6	0	1.31×10^{-23} [a]

[a] significant ($\alpha = 0.05$) for both one- and two-sided p-value

TABLE VI. Mann-Whitney U test results between PGS and PG

N	U	p (one-sided)
10^2	319.5	7.01×10^{-11} [a]
10^3	1	2.74×10^{-18} [a]
10^4	0	5.54×10^{-19} [a]
10^5	0	3.97×10^{-21} [a]
10^6	0	1.31×10^{-23} [a]

[a] significant ($\alpha = 0.05$) for both one- and two-sided p-value

TABLE VII. Mann-Whitney U test results between PGS and Twitter

N	U	p (one-sided)
10^2	1187	3.33×10^{-1}
10^3	758.5	3.29×10^{-4} [a]
10^4	90	1.79×10^{-16} [a]
10^5	0	1.08×10^{-19} [a]
10^6	0	5.17×10^{-23} [a]

[a] significant ($\alpha = 0.05$) for both one- and two-sided p-value

TABLE VIII. Mann-Whitney U test results between PGS and subtitles

N	U	p (one-sided)
10^2	0	3.22×10^{-18} [a]
10^3	0	1.66×10^{-18} [a]
10^4	0	5.30×10^{-20} [a]
10^5	0	1.27×10^{-21} [a]
10^6	0	1.31×10^{-23} [a]

[a] significant ($\alpha = 0.05$) for both one- and two-sided p-value

[1] “Project Gutenberg catalog,” http://www.gutenberg.org/feeds/catalog.rdf.zip (2010).

[2] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, Science **331**, 176 (Jan. 2011), `http://www.sciencemag.org/content/early/2010/12/15/science.1199644.abstract`.

[3] J. Gerber, “pysubtitles,” http://oil21.org/ j/code/pysubtitles/download/ (May 2008).