A corpus study of phonological factors in novel English blends
Annual conference of the CLA (University of Regina)

Michael Dow
Université de Montréal

May 30, 2018
Introduction
“Pussy blends” are becoming more and more commonplace online (popularized by “one thicc bih” meme), for instance *thick pussy > thussy*:

*Jean shorts have to be tight and then i [sic] get a thussy and it’s annoying* (step2myworld, 2018, May 22)

- Irregularity of what a single blend can mean, e.g. *thussy < Thor, Thanos, Margaret Thatcher*, etc.
- Irregularity of what form a single subject will yield, e.g. *Thanos also > thanussy; Margaret Thatcher also > thatchussy*
A pilot study (Dow 2018) of ~100 forms found:

- Contribution of material from first word ($W_1$) increases if longer than $W_2$ (i.e., “pussy”) and if not directly expressed in the meme or thread (= “novel”)

- Equal preference between onset-only and CVC$_1$ contribution of $W_1$ in non-novel forms

- Greater $W_1$ contribution when containing low sonority-high sonority CC juncture (e.g. *Manray > manrussy*) and internal sC clusters (e.g. *toadstool > toadstussy*)

- Internal fricatives and <r> encourage loss of <u> from $W_2$ (e.g. *Ms. Puff > puffsy, Barney > barsy*)
Preliminary findings of an expanded corpus study.

Q: What implicit factors are at play, i.e., what makes a good (or bad) “pussy blend”?

A: $W_1$ contribution increases as a function of source word length (up to final rhyme); $W_2$ contribution remains constant.

Q: Can we consider these words as true blends? If not, what else could they be?

A: These words go against a number of properties of blends. They might instead be considered as mid-clipped compounds.
Meme me up, Scotty
-*(u)*ssy blends & the “one thicc bih” meme

- Appearance & explosion of “one thicc bih” meme (text and Ditty videos, e.g. Fig. 1) in May 2017
- Format: “$x$ is one thicc bih, let me see that $y$”; $x = \text{character or famous personality;}$ $y = \text{blend of } x$ (or related word) and *pussy*

Fig. 1: Babadook > babussy
Source: dcparkers, 06/2017
What’s in a meme?

- Documented -(u)ssy forms date back to early 2010s in gay slang, re-popularized by an April 2017 Tumblr post, in particular:
  - Boy, man > bussy, mussy
  - Throat > thrussy

- thicc & bih AAVE slang (together ≈ “sexy individual”), each documented back as far back as early 2000s
Life cycle of a meme

- Widespread media recognition (e.g. New York Magazine, Buzzfeed) → Ditty app #1 on iTunes store (May 2017)
- Decline around July 2017 (Fig. 2)

Fig. 2: “one thicc bih” in Google trends
Life after thiccnness

- Popularity hard to quantify outside of meme: no substring searches on Twitter
- Several individual searches turn up recent results, though not as common as in summer 2017
- Remains to be seen if the process and/or certain forms survive, but “not dead yet” for the moment
Why study it?

- Memetic nature of “ussification” may resolve some empirical problems in study of blends:
  - Difficulty of automatic collection/recognition (Fradin 2015) → ease of collecting large corpus of meme
  - High degree of variation within and across languages (different “species”) → controlled setting (W2 remains constant) allows for isolation of factors in W1
  - Differing degrees of felicity → several metrics (e.g., meme-user judgments, retweets & likes) can make sense of variation
Will it blend?
Properties of blends

- Definition: “[I]ntentional coinage of a new word by fusing parts of at least two source words of which either one is shortened in the fusion and/or where there is some form of phonemic or graphemic overlap of the source words” (Gries 2004)

- Three salient properties (Fradin 2015)
  - No preservation of lexical integrity: stems are rarely maintained intact & their alteration is variable
  - No fixed pattern of compositionality: head member is unpredictable
  - “Type hapaxes”: blends cannot form series (e.g. *élevage ‘breeding’ + vache ‘cow’ → élévache ‘cow breeding’ but *éléchien ‘dog breeding’)
Overlap

- What determines the respective contribution and order of words in a blend?
- Semantic motivation (*brunch*) vs. phonological selection *glitterati*, cf. Fradin’s (2015) criterion of overlap

<table>
<thead>
<tr>
<th></th>
<th>A. Trunc. = both</th>
<th>B. Trunc. = 1</th>
<th>C. Trunc. = 2</th>
<th>D. Trunc. = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>+OV</td>
<td>daxpór</td>
<td>knáuros</td>
<td>Müillionärin</td>
<td>Paradiesel</td>
</tr>
<tr>
<td>+LIN</td>
<td>daxáf × laxpór</td>
<td>knástos × áuros</td>
<td>Müll × Millionärin</td>
<td>Paradies × Diesel</td>
</tr>
<tr>
<td>+OV</td>
<td>dialügisch</td>
<td>carnivbleous</td>
<td>—</td>
<td>hypocritiquement</td>
</tr>
<tr>
<td>−LIN</td>
<td>dialogisch × Lüge</td>
<td>carnivorous</td>
<td>×</td>
<td>hypocritement × critique</td>
</tr>
<tr>
<td>−OV</td>
<td>brunch</td>
<td>klafúda</td>
<td>smothercate</td>
<td>sálkal</td>
</tr>
<tr>
<td>+LIN</td>
<td>breakfast × lunch</td>
<td>klára × fúda</td>
<td>smother × suffocate</td>
<td>sál × kál</td>
</tr>
<tr>
<td>−OV</td>
<td>agitprop</td>
<td>—</td>
<td>—</td>
<td>rajolivissant</td>
</tr>
<tr>
<td>−LIN</td>
<td>agitation × propag.</td>
<td>—</td>
<td>—</td>
<td>ravissant × joli</td>
</tr>
</tbody>
</table>

Fig. 3: Typology of blends (Fradin 2015)
Chunnel vs. brunch

- Gries’ (2004) Similarity Index (SI), proportionate amount of material contributed by each word:
  - $<ch>a<nnel> + t<unnel> = 0.67$
  - $<br>eakfast + l<unch> = 0.3$
- Average SI of intentional & error-driven blends $\approx 0.5$, vs. random word pairings $\approx 0.3$

Fig. 4: SI by blend type
Extragrammaticality ≠ irregularity

- Debate over blending as morphological (e.g., Bat-El 1996, Plag 2003) vs. extragrammatical (e.g., Bauer 1988, Dressler 2000) process
- Extragrammaticality does not exclude influence of regular/universal linguistic forces, especially phonological for blends (Fradin, Montermini & Plénat 2009)
- Other peripheral (informal) processes evidence knowledge of grammar-external structures or forces, e.g., expletive infixation (McCarthy 1982), shitgibbons (Tessier & Becker 2018)
Blend-trends

- The shorter source word of a blend more likely to contribute more information for intelligibility (Kaunisto 2000)
- Gries’ (2004) results confirm this + a (competing?) tendency for $W_2$ to contribute more:

<table>
<thead>
<tr>
<th>which word is larger?</th>
<th>which word contributes more to the blend?</th>
<th>row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>=</td>
<td>=</td>
<td>122</td>
</tr>
<tr>
<td>source word$_1$</td>
<td>source word$_1$</td>
<td>314</td>
</tr>
<tr>
<td>source word$_2$</td>
<td>source word$_2$</td>
<td>552</td>
</tr>
<tr>
<td>column totals</td>
<td></td>
<td>988</td>
</tr>
</tbody>
</table>

Fig. 5: Contribution by length, phonemes
Methodology
Corpus

- 4450 Tweets scraped using GetOldTweets
- Criteria:
  - June 2017 - August 2017
  - Contains phrases “one thicc bih” & “see that”
- Information automatically gathered:
  - Username
  - Date
  - Retweets & favourites
  - Link, ID
  - Text
Processing

- First 1500 manually annotated for:
  - Full referent
  - Deduced $W_1$ "base" & contribution
  - $W_2$ contribution
  - Standardized blend word (e.g. reduction of $yyyyy$ to $y$)
  - Novelty of blend (if $W_1 \neq$ referent)

- Unclear forms (e.g. inside jokes), tweets using existing words (e.g. Claire de Lune $>$ Debussy), and retweets/identical tweets excluded $\rightarrow$ 1334 forms

- Some educated guesses on bases (e.g., Chuck E. *Cheese* $>$ chussy)
Users & lifespan

- 1156 unique users
- Max no. tweets/user = 6, mean 1.2
- Peak popularity in corpus at end June, declines early in July (Fig. 6, cf. Fig 2)

Fig. 6: No. tweets over time in annotated corpus
More processing

- Generous count of shared graphemes between words, e.g., graphemes in *Bass > bassy*: $W_1 = 4$, $W_2 = 3$
- Number of syllables identified as number of vowels (digraphs & graphemes such as <i> in <-tion> pre-processed), manually verified
Proportion of material each word contributes to the blend (Gries 2004)

Calculation performed on graphemes (\(G = \text{no. graphemes}, \ r = \text{root}, \ c = \text{contribution to blend}, \ b = \text{blend})\), generous interpretation:

\[
\frac{(G_{c1}/G_{r1} \times G_{c1}/G_b) + (G_{c2}/G_{r2} \times G_{c2}/G_b)}{2}
\]

For example, syllabus > syllabussy:

\[
\frac{(8/8 \times 8/10) + (4/5 \times 4/10)}{2} = 0.56
\]
Results
Anecdotes from the data

- /-ɔsi/ final words rare in corpus but highly felicitous; not specially quantified by current measures (e.g. democracy > democrussy)
- Non-contiguous blends also rare but may be of further interest, for instance:
  - octoling chocolate > octochussy
  - hentai Trump > hentrumpussy
- Presence of “intrusive” letters (belonging to neither word) in a handful of forms:
  1. <r> (e.g. Fionn > fiorussy)
  2. <b> (e.g. Jake > jabussy)
  3. <p> (e.g. Kirsten > kirpussy)
  4. C₁ (e.g. Gao > gagussy, Cameron > cacussy, me > memussy)
Preliminary numbers

- Only 130 forms (9.7%) have motivated graphemic overlap (i.e. <u-u> or <p-p> overlap)
- 1187 forms (88.7%) are non-novel
- <ussy> by far the most common W2 contribution (1238), vs. <pussy> (50) and <ssy> (50)
- Only 114 forms (8.5%) have true hiatus at juncture (i.e., <u-u> overlap ignored, as in communism > commussy)
- Mean SI = 0.37 (non-novel), 0.42 (novel)
- Sonority profile of last 2 consonants of CC(C)-final W1 contribution: even (27), rising (84), falling (161)
Syllable size

Preference for 1-syllable $W_1$ contribution over onset-only regardless of novelty:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Onset-only</th>
<th>1σ</th>
<th>2σ</th>
<th>3σ</th>
<th>4σ</th>
<th>5σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-novel</td>
<td>286</td>
<td>719</td>
<td>151</td>
<td>27</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Novel</td>
<td>8</td>
<td>61</td>
<td>63</td>
<td>13</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1: No. forms per $W_1$ syllable contribution by novelty
Expanded syllables, $W_1$

Non-base-final consonant sequences & $<u-u>$ junctures simplified. Max 2 syllables.

<table>
<thead>
<tr>
<th>Shape</th>
<th>Count (Non-novel)</th>
<th>Count (Novel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>208</td>
<td>3</td>
</tr>
<tr>
<td>CC</td>
<td>106</td>
<td>6</td>
</tr>
<tr>
<td>CCC</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>CV</td>
<td>62</td>
<td>3</td>
</tr>
<tr>
<td>CVC</td>
<td>466</td>
<td>26</td>
</tr>
<tr>
<td>CVCC</td>
<td>183</td>
<td>28</td>
</tr>
<tr>
<td>CVCCC</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>CVCV</td>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>CVCVC</td>
<td>63</td>
<td>42</td>
</tr>
<tr>
<td>CVCVCC</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>CVCVCCC</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: $W_1$ contribution syllable shape by novelty
### Results (graphemes)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Gr ($W_1$)</th>
<th>Gr ($W_2$)</th>
<th>Loss</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_1 &lt; W_2$</td>
<td>2.5</td>
<td>4</td>
<td>1.2</td>
<td>235</td>
</tr>
<tr>
<td>$W_1 = W_2$</td>
<td>2.9</td>
<td>4</td>
<td>2.1</td>
<td>318</td>
</tr>
<tr>
<td>$W_1 &gt; W_2$</td>
<td>3.8</td>
<td>4</td>
<td>3.5</td>
<td>634</td>
</tr>
<tr>
<td>$W_1 &lt; W_2$</td>
<td>3.4</td>
<td>4</td>
<td>0.5</td>
<td>23</td>
</tr>
<tr>
<td>$W_1 = W_2$</td>
<td>4.1</td>
<td>4</td>
<td>0.9</td>
<td>33</td>
</tr>
<tr>
<td>$W_1 &gt; W_2$</td>
<td>6</td>
<td>4</td>
<td>1.9</td>
<td>95</td>
</tr>
</tbody>
</table>

**Table 3:** Mean grapheme (Gr) contribution & loss by novelty and relative length
Results (syllables)

Fig. 7: No. syllables, source word vs. contribution ($W_1$)
Discussion & future work
Findings

- Grapheme contribution of $W_1$ increases with source word length regardless of novelty, while $W_2$ contribution stays the same.
- Novel forms on average lose less information.
- Competing preference between $C(VC)$ and “all but final rhyme” templates.
- Loss of $<u>$ from $W_2$ is rare and currently not predictable by any factor.
Are they blends?

- *(u)*ssy forms meet some criteria, but not series-like nature & tendency for $W_1$ to contribute more as it gets longer, esp. in novel forms

- Many forms are dependent, not transparency of form, for meaning

- Forced combination regardless of overlap: `<chick>en + p<u ssy> → chickussy` (more common) vs. `<Bloss>om + pu<u ssy> → blossy`

- Low SI, especially in non-novel forms

- May in fact be median-clipped compounds (Tournier 1985, Jamet 2009), e.g. *smoke fog > smog*, though not always distinguished from blends
Future work

- Finalization of corpus annotation
- Integration of stress & phoneme counts (size, contribution, better sonority profiles) into results
- Comparison with fandom pairing names (DiGirolamo 2012), especially stress match (location of juncture) & onset conservation (e.g. *Clyde + Rani > Clani*)
- Judgment task (variation & strength of factors)
Thank you!
Acknowledgments

- The makers of the Ditty app
- Darius Kazemi (maker of the UssyBot)
- The audience members of the MOT 2018 (including Jeffrey Lamontagne & Daniel Curie Hall)


