An Expanded Taxonomy of Semiotic Classes for Text Normalization

Daan van Esch\(^1\), Richard Sproat\(^2\)

\(^1\)Google, Mountain View, CA, USA
\(^2\)Google, New York, NY, USA

dvanesch@google.com, rws@google.com

Abstract

We describe an expanded taxonomy of semiotic classes for text normalization, building upon the work in [1]. We add a large number of categories of non-standard words (NSWs) that we believe a robust real-world text normalization system will have to be able to process. Our new categories are based upon empirical findings encountered while building text normalization systems across many languages, for both speech recognition and speech synthesis purposes. We believe our new taxonomy is useful both for ensuring high coverage when writing manual grammars, as well as for eliciting training data to build machine learning-based text normalization systems.

Index Terms: text normalization, taxonomy, non-standard words, semiotic classes

1. Introduction

Automatic speech recognition (ASR) and text-to-speech (TTS) systems typically require comprehensive language-specific text normalization processing pipelines to handle tokens like numbers (“103”), URLs (“www.foo.bar”), and so on. Sproat et al. [1] provided the first systematic overview of such Non-Standard Words — NSWs, noting that normalization of NSWs is a frequently overlooked, yet critical area in building ASR and TTS systems alike.

Over the years, we have collected many additional types of NSWs that one will encounter when building a general-purpose speech processing system that aims to be robust to different domains and registers. In this paper, we present a categorization of the types of NSWs we are currently aware of, along with commentary where needed. We also describe how we used our taxonomy to accelerate development of text normalization systems for new languages by allowing us to systematically elicit linguistic knowledge from experts using our taxonomy.

We hope that this paper can help others speed development of new text normalization systems and improve coverage across NSW categories. We also hope to highlight once again the complexity of text normalization and the fact that though deep learning is revolutionizing ASR and TTS in general, there are still areas requiring careful application of linguistic knowledge. The taxonomy in this paper can be used to create templates for manual grammars, but it can equally be used to structure data collections aimed at training machine learning models.

In this paper we use the terms “verbalize” or “verbalization” as an expression that is hopefully neutral to whether we are concerned primarily with ASR or TTS: thus when we speak of the “verbalization” of a token, we mean that a TTS system would read it in a given way, or that an ASR system would expect a speaker to say it in a particular way.

2. The Taxonomy of [1]

We start by presenting the taxonomy of NSWs from [1] in Table 1. The taxonomy classifies NSWs into three broad categories: those that are largely alphabetic, those involving numbers, and miscellaneous instances, at least some of which do not fall neatly into either category. Within each category, a taxonomy is given that partly depends on the kind of operation that is involved in mapping from the input token to its verbalization; and partly on functional considerations of how the token is used.

Looked at from today’s perspective, there are obvious omissions from this set. Some of these reflect types of expressions that simply did not occur in the late 1990’s and early 2000’s, such as Twitter hashtag tokens. Others were less widely represented: thus “funny spellings” such as $\text{sllooooww}$ occurred in SMS, but it was not until the advent of Twitter and other social media that such uses became so prominent, engendering a small cottage industry of work on text normalization aimed at such examples [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

Finally, there are many categories that were omitted simply because they did not arise much or at all in the types of data considered in the project at the time. As TTS and ASR systems have become significantly more widely used over the past fifteen years, new categories of NSWs have come into the fore. Frequently these kinds of examples show up in practice when a TTS system, say, gets them wrong and a user submits feedback.

3. Non-Standard Words: An Expanded Taxonomy

In this section we present, in a series of tables, a more complete set of NSW types, broken down into a set of twelve broad categories. Included here are the original categories in [1], indicated in upper-case boldface. For ease of presentation to an international audience, and to obviate the need to provide glosses, we represent examples from English. It is of course to be understood that many or all of these types will occur in most languages.

Wordlike Tokens. Many of the categories that are broadly “wordlike” in that they are written mostly using the normal alphabet for the language in question have already been described in [1]. See Table 2. The main categories that we add here are CENSORED, as used in particular for off-color terms that are censored in writing, and for which a speech system must make some decision on how they are to be verbalized. 1337-speak (leet-speak) is a special category where digits are used to represent similar looking letters; thus even though the actual symbols used are numerical symbols, they are being used as surrogates for alphabetical symbols. AMPERSAND-WORD is really an instance of letter sequence (LSSEQ), but includes the non-letter symbol ‘&’, in English conventionally verbalized as and. Along similar lines, musical notes contain letters coupled with domain-specific symbols. Finally TRANSLIT refers, with slight abuse of terminology, to words that appear in a foreign...
Table 1: The taxonomy of non-standard words as presented in [1].

<table>
<thead>
<tr>
<th>EXPN</th>
<th>LSEQ</th>
<th>ASWD</th>
<th>MSPL</th>
<th>NUM</th>
<th>NORD</th>
<th>NTEL</th>
<th>NDIG</th>
<th>NIDE</th>
<th>NADDR</th>
<th>M</th>
<th>B</th>
<th>E</th>
<th>R</th>
<th>S</th>
<th>U</th>
<th>NONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>abbreviation</td>
<td>letter sequence</td>
<td>read as word</td>
<td>misspelling</td>
<td>number (cardinal)</td>
<td>number (ordinal)</td>
<td>telephone (or part of)</td>
<td>number as digits</td>
<td>identifier</td>
<td>number as street address</td>
<td>zip code or PO Box</td>
<td>a (compound) time</td>
<td>a (compound) date</td>
<td>year(s)</td>
<td>money (US or other)</td>
<td>money tr/millions</td>
</tr>
<tr>
<td>adv, N.Y, mph, govt</td>
<td>CIA, D.C., CDs</td>
<td>CAT, proper names</td>
<td>geography</td>
<td>12, 45, 1/2, 0-6</td>
<td>May 7, 3rd, Bill Gates III</td>
<td>212 555-4523</td>
<td>Room 101</td>
<td>747, 386, I5, pc10, 3A</td>
<td>5000 Pennsylvania, 4523 Forbes</td>
<td>91020</td>
<td>3:20, 11:45</td>
<td>2/2/99, 14/03/87 (or US) 03/14/87</td>
<td>1998, 80x, 1900c, 2003</td>
<td>$3-45, HKS300, Y20,000, $200K</td>
<td>$3-45 billion</td>
<td>75%, 3.4%</td>
</tr>
</tbody>
</table>

Table 2: Word-like tokens.

<table>
<thead>
<tr>
<th>EXPN</th>
<th>LSEQ</th>
<th>ASWD</th>
<th>MSPL</th>
<th>NUM</th>
<th>NORD</th>
<th>NTEL</th>
<th>NDIG</th>
<th>NIDE</th>
<th>NADDR</th>
<th>M</th>
<th>B</th>
<th>E</th>
<th>R</th>
<th>S</th>
<th>U</th>
<th>NONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>letter sequence</td>
<td>misspelling</td>
<td>number (cardinal)</td>
<td>number (ordinal)</td>
<td>telephone (or part of)</td>
<td>number as digits</td>
<td>identifier</td>
<td>number as street address</td>
<td>zip code or PO Box</td>
<td>a (compound) time</td>
<td>a (compound) date</td>
<td>year(s)</td>
<td>money (US or other)</td>
<td>money tr/millions</td>
<td>percentage</td>
<td>mixed or “split”</td>
</tr>
<tr>
<td>adv, N.Y, mph, govt</td>
<td>CIA, D.C., CDs</td>
<td>geography</td>
<td>12, 45, 1/2, 0-6</td>
<td>May 7, 3rd, Bill Gates III</td>
<td>212 555-4523</td>
<td>Room 101</td>
<td>747, 386, I5, pc10, 3A</td>
<td>5000 Pennsylvania, 4523 Forbes</td>
<td>91020</td>
<td>3:20, 11:45</td>
<td>2/2/99, 14/03/87 (or US) 03/14/87</td>
<td>1998, 80x, 1900c, 2003</td>
<td>$3-45, HKS300, Y20,000, $200K</td>
<td>$3-45 billion</td>
<td>75%, 3.4%</td>
<td>WS99, x220, 2_car</td>
</tr>
</tbody>
</table>

Table 3: Basic numbers.

<table>
<thead>
<tr>
<th>NUM</th>
<th>NORD</th>
<th>NDIG</th>
<th>DECIMAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>cardinal numbers</td>
<td>ordinal numbers</td>
<td>number as digits</td>
<td>1.34, -1.34</td>
</tr>
<tr>
<td>12, 45, -5, Superbowl XLVIII</td>
<td>May 7, 3rd, Bill Gates III</td>
<td>Room 101,</td>
<td></td>
</tr>
</tbody>
</table>

script, such as the word Facebook appearing in Roman letters in an otherwise Cyrillic Russian text. Generally, such words are handled by first transliterating them to the target script, and then pronouncing them as a normal word of the language. **Basic Numbers.** Basic numbers are given in Table 3. Of these, only DECIMAL numbers are added here which, for whatever reason, were not included as a separate category in the earlier taxonomy.

**Identifiers.** Identifiers, given in Table 4 are a diverse class of mostly numerical tokens used to identify specific entities. Some of the categories listed are rather specific, such as runways, reflecting the fact that these may have rather context-specific verbalizations that one just has to know in order to verbalize them.

The fine-grainedness of the categories reflects the fact that in some languages one may need to know that something belongs to that category in order to know that it must be verbalized in a specific way. For example, sports club names involving a name plus the year in which the club was formed are actually not special in English: Astana 1964 would be verbalized as Astana nineteen sixty four, the only requirement being the detection of 1964 as a year. However, in Japanese and Korean, 1964 in such cases would be verbalized as a digit sequence, rather than as a year (which in both languages would be verbalized as an ordinary cardinal number).

The two main categories that are not primarily numeric are URL and a new category MENTION, a device that includes hashtags and ampersand-prefixed mentions of handles, which have become much more prevalent with the advent of Twitter. Note that both MENTION and URL may contain contiguous strings of letters that need to be broken down further — rickysync.com as ricky’s n y c dot com — or that require further normalization — 0th.org as zeroeth dot org.

**Dates, Times.** Examples of dates and times can be found in Table 5.

**Ratios, Percentages, Fractions.** Percentages, ratios and fractions, which are mostly self-explanatory, are given in Table 6. Ratings are an interesting category, in that they look somewhat like fractions, but often have a quite different verbalization. We put star ratings (“****”) here even though they are not numeric in form.

**Geographic.** Geographic entities are presented in Table 7. Again, many of these are self-explanatory though some of them present some interesting complexities. For example exit designations on highways require some care: AZ/W141 should be...
Wire gauge verbalizations are particularly idiosyncratic.1 Table 9 gives various sports-related instances. Verbalization of “0” in tennis is a well-known idiosyncratic case where the letters W and E are verbalized as double u and e, not as one would have expected as west and east.

Measures. Various categories of measures are given in Table 8. Of these, compound expression such as dimensions can be tricky, so that for example 1024x768 should be verbalized ten twenty four by seven sixty eight, and not for examples as 1024x768 is verbalized as ten twenty four by seven sixty eight, or other conceivable verbalizations. Wire gauge verbalizations are particularly idiosyncratic.1

Sports. Table 9 gives various sports-related instances. Verbalization of “0” in tennis is a well-known idiosyncratic case in English, but other sports include idiosyncratic notations and verbalizations. Thus Nc6 in chess would be knight to c six.

Currency. Currency expressions from [1] are given in Table 10. One issue to note with English currency and measure expressions is the verbalization of plurals in compounds like five dollar bill, which is five dollar bill, not five dollars bill.

Formulae. Table 11 contains examples of mathematical and chemical formulae, both of which require special treatment. Note here that we are primarily interested in relatively simple cases that might occur inline in text like $\sqrt{3}$ verbalized as square root of three, or PtF6 verbalized as platinum hexafluoride. Very complex expressions go beyond the scope of text normalization. For example, complex mathematical formulae are treated in a classic system called Aster [16].

Non-Linguistic. Various non-linguistic or semi-linguistic symbols arise that pose text normalization problems. These include:

• Repetition symbols which designate that the preceding material is to be reduplicated. In some languages, these can be particularly challenging, e.g. in Thai, which does not mark word or phrase boundaries in text, meaning that the system must compute exactly how much of the preceding material to reduplicate.

• Elision symbols.

• Emoticons and emoji: these are typically easy to identify, but beyond the most common cases, it is often hard to know how one should verbalize them.

• Shhhhh, and other interjections: Then he was all like, “yeah huh”’ And I was all like, “nuh uh” And he was like “HAH!” and so I gave him the stink eye. These often involve non-speech sounds, or speech sounds that are used in a way not conformant to the phonotactics of the language (shhhhh is not a phonotactically well-formed word in English, for example).

No Verbalization. Table 12 gives two categories where the tokens are unverbalized, both already presented in [1]. The categories here, SLNT and NONE, are treated differently. Neither are verbalized as words, but SLNT tokens such as the “/” boundary marker in large kit./3BR/2BA may have prosodic effects insofar as they often correspond to phrase boundaries. In contrast, those in the NONE category would typically be ignored (though in a TTS application one might want to indicate that there some material that is not being read, cf. [17]).

4. Accelerating Development with the Taxonomy

We have used the taxonomy above to collect text normalization data from speakers across dozens of languages. In these collections, we typically show native speakers — who are usually...
Table 7: Geographic entities.
NADDR building numbers 5000 Pennsylvania, 4523 Forbes
NZIP zip code 91020, 23945-2345, 1039 AA, 2
PO Box PO Box 1
Lat-long 52°22′N 4°54′E
Street numbers W 17th St, Hougang Street 21, 14 St, 8 Ave
Within-building numbers #301, apt 301
Area/grid numbers 1200 South, Pier 39
Exit numbers Exit 314 towards AZ/W141
Provinces/states/countries OH, CA, MA, QLD, NB, USA, UK, MX, BR
Highway numbers I-280, S101, A113, Route 66, CA 17, US 101, I35W

Table 8: Measure expressions.
Measures km, 3 mi, ft-lbs
Square measures km², 10 acres, 10 ha
Cubic measures 30 cu ft, 30 m³
Relative measures km/h, mph, 4,034/km², 234 mpg, 10 l/100km
Meas. with punct. 6°, 5°8′
Temperatures 0 C, 5°, -5°
Dimensions 1024x768, 10x4x8
Stock indices . . . opened 17,652.36 (insert “points”)
Wire gauge 1/0 (one aught), 2/0 (two aught)

Table 9: Sports-related expressions.
Plain scores 3-1 three to one
Tennis 15-0 (fifteen love)
Australian football 10.12 (72) (ten twelve seventy two)
Chess notation Nc6, Rxc6

Table 10: Currency expressions.
MONEY — money $3.45, HK$300, Y20,000
BMONEY — money tr/m/billions $3.45 billion

Table 11: Formulae.
Mathematical formulae \(5/6, \sqrt{3}\)
Chemical formulae CsCl, H2SO4, PtF6

Table 12: No Verbalization.
SLNT — silent boundary or emphasis character:
large kit./3BR/2BA, KENT*RLTY, **YES**
NONE — ignored “ASCII art”, formatting junk,
unknown Unicode characters

not trained computational linguists — a number of examples from each category, and we ask them to enter the appropriate verbalizations. We also allow our consultants to skip if they are unsure. Finally, every category allows for free-form text input so responders can add further clarification as necessary (e.g. when two variants are possible). We find that linguistically aware native speakers are typically able to complete these questionnaires in just a few hours of work.

Once we have responses to our questionnaire, we use basic scripts to infer some basic verbalization rules from the responses, e.g. in “#twitter”, we determine what the verbalization for “#” is based on the verbalization entered by the consultant (in this case, hashtag in English). Then, we insert these newly gathered values into our Thrax-based text normalization grammars [18, 19, 20]. Other, more complex examples that we gather are added as unit tests so our in-house linguists can quickly iterate on the grammars to achieve the correct output. All of this improves development turn-around times by days, especially for languages without in-house linguistic expertise.

Of course, this taxonomy can also be used to ensure high coverage is achieved in collecting data for machine-learning text normalization systems. For easy cases like “#twitter”, it remains more cost-effective to use rule-based grammar. Clearly, though, machine learning becomes more and more attractive as one ventures deeper into this taxonomy, e.g. years in sports club names as discussed above require an idiosyncratic set of rules, where a machine learning system might generalize better.

5. Conclusions
We have presented an expanded taxonomy of NSWs that will be useful for developers building text normalization systems for modern-day, open-domain ASR or TTS applications. Any text normalization system that hopes to be robust to a wide diversity of inputs must cover a great number of idiosyncratic cases. Our taxonomy organizes a large number of cases we have run into over the years. We will no doubt encounter others: text normalization is a veritable rabbit hole of corner cases.

In the age of deep learning, verbalizing NSWs remains an area where linguistic knowledge is virtually always injected explicitly, whether it be through rule-based grammars or carefully designed data for machine learning systems. We hope our expanded taxonomy will be of use to developers, both to increase coverage, as well as to accelerate development.

6. Acknowledgements
We would like to thank all participants in the 2016 Google Text Normalization summit in New York for contributing various examples during a particularly productive break-out session.
7. References


4020