INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

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* — **NSW**s, 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 *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.

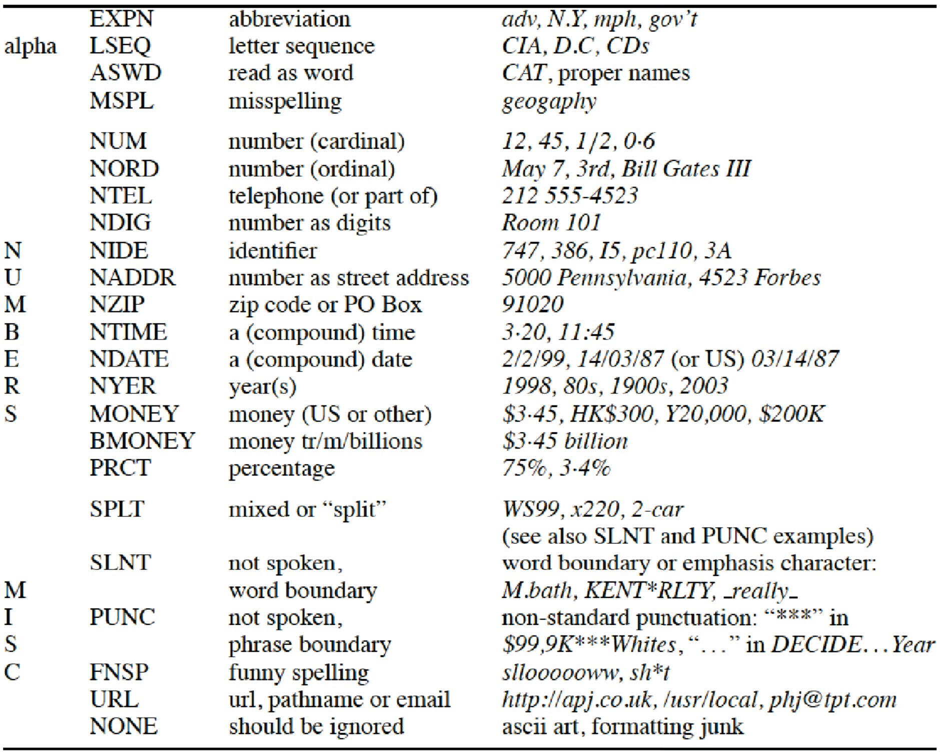
3. Non-Standard Words: An Expanded Taxonomy

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 restrict examples to 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 (LSEQ), 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

http://dx.doi.org/10.21437/Interspeech.2017-402

Table 1: *The taxonomy of* non-standard words *as presented in [1].*



	EXPN	abbreviation	*adv, N.Y, mph, gov't*
alpha	LSEQ	letter sequence	*CIA, D.C, CDs*
	ASWD	read as word	*CAT*, proper names
	MSPL	misspelling	*geogaphy*
	NUM	number (cardinal)	*12, 45, 1/2, 0·6*
	NORD	number (ordinal)	*May 7, 3rd, Bill Gates III*
	NTEL	telephone (or part of)	*212 555-4523*
	NDIG	number as digits	*Room 101*
N	NIDE	identifier	*747, 386, I5, pc110, 3A*
U	NADDR	number as street address	*5000 Pennsylvania, 4523 Forbes*
M	NZIP	zip code or PO Box	*91020*
B	NTIME	a (compound) time	*3·20, 11:45*
E	NDATE	a (compound) date	*2/2/99, 14/03/87* (or US) *03/14/87*
R	NYER	year(s)	*1998, 80s, 1900s, 2003*
S	MONEY	money (US or other)	*$3·45, HK$300, Y20,000, $200K*
	BMONEY	money tr/m/billions	*$3·45 billion*
	PRCT	percentage	*75%, 3·4%*
	SPLT	mixed or "split"	*WS99, x220, 2-car*
			(see also SLNT and PUNC examples)
	SLNT	not spoken,	word boundary or emphasis character:
M		word boundary	*M.bath, KENT*RLTY, _really_*
I	PUNC	not spoken,	non-standard punctuation: "***" in
S		phrase boundary	*$99,9K***Whites*, "..." in *DECIDE...Year*
C	FNSP	funny spelling	*slloooooww, sh*t*
	URL	url, pathname or email	*http://apj.co.uk, /usr/local, phj@tpt.com*
	NONE	should be ignored	ascii art, formatting junk

Table 2: *Word-like tokens.*

EXPN — abbreviation	adv, N.Y, mph, govt
LSEQ — letter sequence	CIA, D.C., CDs
MSPL — misspelling	geogaphy
FNSP — funny spelling	sllooooww, cul8r
CENSORED	sh*t, f***
ASWD — verbalized as word	CAT, NATO, LASER, GOOG
1337-speak	1337, n00b
SPLT	C-SPAN
AMPERSAND-WORD	AT&T, H&M
MUSIC-NOTE	C, D, Re
TRANSLIT	words in another script

Table 3: *Basic numbers.*

NUM — cardinal numbers	12, 45, -5, Superbowl XLVIII
NORD — ordinal numbers	May 7, 3rd, Bill Gates III
NDIG — number as digits	Room 101,
DECIMAL	1.34, -1.34

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 — *rickysnyc.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

Table 4: *Identifiers.*

NIDE — identifier	747, 386, I5, pc110, 3A U, A4, 401(k)'s
NTEL — telephone	212 555-4523
Version numbers	Android 3.1, 5.2.13, 5.3.alpha
Years in sports club names	Astana 1964
Credit card numbers	1243 4567 8910 1234
PIN numbers, CCV	1234
Social Security Numbers	123-45-6789
ID numbers	S123456789
Seasons/episodes	S02E02
Flight numbers	AAL12, Flight 637
Call signs	808BL (*eight oh eight bravo lima*)
Runways	13L / 13R (*thirteen left*, *thirteen right*)
URL — url, pathname or email	http://apj.co.uk, /usr/local, phj@tpt.com, foobar.jpg, test1.js
MENTION	#hashtag, @person, +person

Table 5: *Dates and times.*

NTIME — a (compound) time	3:20 p.m., 11:45, 15:38, 3 o'clock, 3pm, 2:02:57, 8ish
NDATE — a (compound) date	2/2/99, 14/03/87 (or US) 03/14/87, March 3rd 2019, Mar 3, 6/87, 31.VII.1932
NYER — year(s)	1998, 80s, 1900s, 2003, '70s, John Smith '19, 29 BCE,
Durations	3hr30m, 4h2m23s, 2:02:57, 24/7/365
Time zones	GMT, UTC+1, UTC+5:45

Table 6: *Ratios, percentages and fractions.*

PRCT — percentage	75%, 3.4%, -3%
Fractions	$\frac{1}{3}$, 1/3 , 1 $\frac{1}{2}$, 1 1/2
Ratios	2:3, 3.37:1
Ratings	4.5/5, **** (*four stars*)

verbalized as *Arizona west one forty one*. Highway numbers can also present interesting corner cases, such as *I35W/I35E*, which are branches of I35 around Minneapolis and again in Dallas, 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 *one oh two four eks seven six 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 *$5 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 hexaflouride*. 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

[1] https://en.wikipedia.org/wiki/American_wire_gauge#Pronunciation

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	km^2, 10 acres, 10 ha
Cubic measures	30 cu ft, 30 m^3
Relative measures	km/h, mph, 4,034/km^2, 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 ^ 3`, $\sqrt{3}$
Chemical formulae	CsCl, H2S04, 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 respondents 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

[1] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Computer Speech and Language*, vol. 15, no. 3, pp. 287–333, Jul. 2001. [Online]. Available: http://dx.doi.org/10.1006/csla.2001.0169

[2] Y. Xia, K.-F. Wong, and W. Li, "A phonetic-based approach to Chinese chat text normalization," in *ACL*, Sydney, Australia, 2006, pp. 993–1000.

[3] M. Choudhury, R. Saraf, V. Jain, S. Sarkar, and A. Basu, "Investigation and modeling of the structure of texting language." *International Journal of Document Analysis and Recognition*, vol. 10, pp. 157–174, 2007.

[4] C. Kobus, F. Yvon, and G. Damnati, "Normalizing SMS: are two metaphors better than one?" in *COLING*, Manchester, UK, 2008, pp. 441–448.

[5] R. Beaufort, S. Roekhaut, L.-A. Cougnon, and C. Fairon, "A hybrid rule/model-based finite-state framework for normalizing SMS messages," in *ACL*, Uppsala, Sweden, 2010, pp. 770–779.

[6] M. Kaufmann, "Syntactic normalization of Twitter messages," in *International Conference on NLP*, 2010.

[7] F. Liu, F. Weng, B. Wang, and Y. Liu, "Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision," in *ACL*, Portland, Oregon, USA, 2011, pp. 71–76.

[8] D. Pennell and Y. Liu, "A character-level machine translation approach for normalization of SMS abbreviations." in *IJCNLP*, 2011.

[9] A. T. Aw and L. H. Lee, "Personalized normalization for a multilingual chat system," in *ACL*, Jeju Island, Korea, 2012, pp. 31–36.

[10] F. Liu, F. Weng, and X. Jiang, "A broad-coverage normalization system for social media language," in *ACL*. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 1035–1044.

[11] X. Liu, M. Zhou, X. Zhou, Z. Fu, and F. Wei, "Joint inference of named entity recognition and normalization for tweets," in *ACL*, Jeju Island, Korea, 2012, pp. 526–535.

[12] H. Hassan and A. Menezes, "Social text normalization using contextual graph random walks," in *ACL*, 2013, pp. 1577–1586.

[13] Y. Yang and J. Eisenstein, "A log-linear model for unsupervised text normalization," in *EMNLP*, 2013, pp. 61–72.

[14] G. Chrupaa, "Normalizing tweets with edit scripts and recurrent neural embeddings," in *ACL*, Singapore, 2014.

[15] W. Min and B. Mott, "NCSU_SAS_WOOKHEE: A deep contextual long-short term memory model for text normalization," in *WNUT*, 2015.

[16] T. Raman, "Audio system for technical readings," Ph.D. dissertation, Cornell University, 1994.

[17] R. Sproat, J. Hu, and H. Chen, "Emu: an e-mail preprocessor for text-to-speech," in *Multimedia Signal Processing*, 1998.

[18] B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, and T. Tai, "The OpenGrm open-source finite-state grammar software libraries," in *ACL*, 2012, pp. 61–66.

[19] H. Sak, F. Beaufays, K. Nakajima, and C. Allauzen, "Language model verbalization for automatic speech recognition," in *Acoustics, Speech and Signal Processing, 2013. ICASSP 2013. IEEE International Conference on*, 2013. [Online]. Available: http://goo.gl/xmCOR

[20] P. Ebden and R. Sproat, "The Kestrel TTS text normalization system," *Natural Language Engineering*, vol. 21, no. 3, pp. 333–353, May 2015. [Online]. Available: https://www.cambridge.org/core/article/div-class-title-kestrel-tts-text-normalization-system-div/F0C18A3F596B75D83B75C479E23795DA