RoCS-MT Normalisation guidelines

Sentence Segmentation

Selected texts are manually segmented into "sentences" (using the segmentation marker |||). This will often correspond to standard end of sentence markers but not always. The aim of this is to create segments of a more standard nature (i.e. corresponding to traditional definitions of a sentence) and also to reduce the length of the sequences in cases where posts are simply a long stream of text. When dealing with non-standard texts, the notion of a sentence is not always straightforward, as texts do not correspond to traditional definitions of a sentence. We nevertheless adhere as best as possible to the definition of a sentence being a clause containing a subject and a predicate.

Example of sentence segmentation in a post lacking punctuation indicating sentence boundaries:
Imo lily for the most part is a rather likeable character (keep in mind ive only played a bit of season 1 up to like part of ep 3)|||unlike her father lily tries to work with lee and for the most part the group |||she does what she has to |||she is a bit of a bitch cause she has to |||i only have 1 problem with lily |||she killed carly |||if you could get ben to confess |||yes i got spoiled on that and a few other things |||ben is useless |||he is what looks like a collage student yet he is useless and needs rescuing 24/7 |||pls dont downvote me into oblivion if i missed something

Example of sentence segmentation for a very long sentence (before *so* and *and*):
And I offered to go out to look at Christmas lights but he was out of town |||so we couldn't do that- and now he's back in town and I asked him what he's doing on New Years and he said he was going to hang out with two of the managers form our work |||and likeeeee he didn't offer an invite and then he said if he cancels with them would I like to come over and I just said idk what I'm doing yet so I don't know\n|||Am I overthinking this? \n|||Does he not care?

Note traditional sentence-final punctuation can be used for emphasis to delimit words that would traditionally form a single sentence (in French this is called *ponctuation forte abusive*). In these cases, we choose to retain the punctuation during normalisation (see below), but keep the sequence of punctuation-delimited text as a single segment when segmenting, for example:

Ok, it's over now. Hopefully.

Normalisation

Removing and cleaning "sentences"

We remove units of text post-sentence segmentation that contain only symbols or non-sensical text. We also strip sentences of surrounding whitespace, including newlines. Any newlines within sentences are removed.

Punctuation

- Simplify repeated punctuation to a single character - ??? > ?, !!!!!! > !, ?!!! > ?!
- Final punctuation:
  - Add any missing final punctuation. The only exception is when there are emojis at the end of the sentence; we do not add additional final punctuation after a final emoji, as we consider that they play the role of punctuation, although we do not remove final punctuation if included after or before the emoji in the original post. By default, the final punctuation added is:
    - Full stop for declaratives
    - ? for questions
    - ! if there is an indicating of excitement/shock etc. (E.g. Gotchaaa -> Gotcha!)
  - Remove question marks when the sentence is not a question (a common feature in English).
- Commas:
  - Add commas where appropriate to make sentences grammatically correct, e.g. to separate out finite clauses, e.g.
    - **Original**: *I was trying to go to sleep bc I couldn't when I got two notifications from this girl I really liked but I didn't look at my phone and I went back to sleep.*
    - **Normalised**: *I was trying to go to sleep, because I couldn't, when I got two notifications from this girl I really liked, but I didn't look at my phone and I went back to sleep.*
  - Final and initial acronyms of surprise (e.g.lol, WTF, etc) to be separated by a comma, e.g.
    - **Original**: *whats going on lol.*
    - **Normalised**: *What's going on, lol?*
- Dot dot dot (…) can be used for different purposes:
  - We retain … in the case of incomplete sentences, reformulations within a sentence or separating clauses in a sentence that would otherwise be ungrammatical.
  - We remove them when they represent pauses in an otherwise grammatical sentence.

End punctuation for emphasis (Ponctuation forte abusive)

- As described previously, we allow end punctuation (i.e. full stops) for emphasis in structures where the first clause is a complete sentence and the second element is a modifier, following the definition in https://aclanthology.org/2010.jeptalnrecital-court.5/, e.g.
  - **Original**: ok it's over now\nhopefully
  - **Normalised**: Ok, it's over now. Hopefully.

Interjections

- Normalise interjections containing repeated characters to minimal versions. In most cases, this results in single rather than repeated characters (ahhh > ah, uhhh > uh), but in some cases requires retaining at least one repeating character (hmmm > hmm, because hm is a different interjection).

Placeholders

- The original posts sometimes contain placeholders where the authors have anonymised elements or are simply giving examples. These placeholders are to be retained during normalisation, with the two possibilities being <*name*> or *[name]*. Variants such as *(name)* should be replaced with the *[name]* placeholder.

Grammatical correction (including elision (e.g. from telegraphic speech, pro-drop and auxiliary drop))

- Grammatical correction should be applied where it involves changing the inflection of a word or adding missing words (i.e. no major changes that significantly change the structure of the sentence). Examples of grammatical correction to be changed:
  - There's <PLURAL NOUN> > There are <PLURAL NOUN>
  - Wrong inflections, e.g. it do > it does, these thing > these things OR this thing (depending on the context)
  - Sometimes "that" should be added, but only when the sentence would otherwise be ungrammatical.
  - Some sentences exhibit characters of telegraphic speech (e.g. dropped pronouns, articles, verbs). The normalised forms reintroduce those words, including to make the sentence grammatical.
    - Missing articles: e.g. Vacuum won't stop sucking > The vacuum won't stop sucking, Going to the park 'I'm going to the park')
    - Missing pronouns, e.g. Going to the park > I'm going to the park. (the pronoun to be added will depend on the context).
    - Missing auxiliaries, e.g. you want this? > Do you want this?

Spelling correction

- Correct all misspellings, except if the spelling error is important to maintain the meaning of the post (i.e. if there is a meta-linguistic question being asked)
- Spelling variants linked to different varieties of English (e.g. BrEn colour vs. AmEn color) are not normalised, i.e. they are kept in their original form.
- Capitalisation: capitals should be normalised (i.e. removing capitals for expressive reasons or when erroneous), adding capitals to named entities where relevant and capital letters at the beginning of sentences.

Truncations/simplifications

- Truncated words are normalised to their full form as long as the long form has the same meaning and is used in natural speech (e.g. pups 'puppies', sub 'subreddit', veggies 'vegetables')
- Truncations includes: words that are truncated (e.g. bro 'brother') and abbreviations (e.g. hr 'hour', w/ 'with', b/c 'because').

- Note that during truncation, words can adopt additional suffixes such as 'y' (e.g. veggies 'vegetables', beardie 'bearded dragon), 'o' (e.g. combo 'combination', convo 'conversation') and 's' (e.g. probs 'probably').

Acronyms

- The expansion or acronyms is dependent on the type and how lexicalised they are.
  - The following acronyms are to be kept as they are (i.e. not expanded):
    - Commonly used, lexicalised acronyms for named entities such as places, universities, countries, institutions, concepts especially if they are pronounced as an acronym (UPLB, DNA, USA, NATO).
    - Acronyms that are best known in their acronym form and would be unnatural in their expanded form (e.g. TLDR, PC, OVR, REP, lol, lmao rofl, smh)
  - Otherwise, acronyms should be expanded, e.g. NYE > New Year's Eve, NAE > North America and Europe, PT > Physical Therapy, IRL > in real life OR real-life (depending on the context). The lexicalisation of acronyms is a continuum, so it may be difficult to decide whether to expand acronyms or not.
- N.B. Acronyms can apply to sequences of words (e.g. fyi 'for your information') but also to sequences of syllables (e.g. nvm 'never mind'). They can also apply partially only to part of a word (e.g. bday 'birthday', gfriend 'girlfriend')

Foreign words

- Foreign words are replaced by their English translation (e.g. Sans guac 'without guacamole') if they are not sufficiently lexicalised.

Elision (word dropping)

- Some sentences exhibit characters of telegraphic speech (e.g. dropped pronouns, articles, verbs). The normalised forms reintroduce those words, including to make the sentence grammatical (e.g. Vacuum won't stop sucking > The vacuum won't stop sucking, Going to the park 'I'm going to the park')
- CHECK adjectival and nominal phrases!!

Phonetically inspired spelling

- Words that use non-orthodox spelling inspired by phonetically similar sequences of characters are normalised to their standard spelling.
- This can include simplification of spelling (e.g. wot 'what', becoz 'because'), digit phonetisation (e.g. 2day 'today', b4 'before'), letter phonetisation (e.g. c u 'see you', u r 'you are'), elision of unaccented syllables (e.g. 'cause 'because', 'bout 'about', nana 'banana'), imitation of dialects or accents (e.g. lil 'little', aight 'alright')
- We make a distinction between mis-spellings and phonetically inspired spellings. The context surrounding the word can often indicate which one applies in case of apparent ambiguity. Phonetically inspired spellings are often obviously so (e.g. digit phonetisation), are different in an exaggerated way that could not plausibly be a spelling error, or in some cases are surrounded by so many other words of the same nature that it is clear that the variations were done on purpose.

Meta-linguistic texts

- When posters are talking about language and using examples, the original words being discussed should be kept largely as they are (e.g. keep spelling errors if that is what is being discussed, keep markers of rhythm if that is important, keep foreign terms if necessary).
- The words should be normalised if the non-standard characteristics are not what is under discussion (e.g. spelling errors can be corrected if the meaning of the post does not change on correction).
- An exception is for repeated characters, which are simplified to their minimal form that shows repetition (e.g. looooooooool becomes loool).

Numerals and symbols

- Digits and letters used to indicate numbers are to be kept as they are (i.e. no homogenisation of 10 and ten)
- The placement of symbols is corrected where it is non-standard (e.g. in English, the currency symbol should go before the number rather than after, as in £40 rather than 40£)
- Other symbols replacing words are normalised (e.g. & 'and', + 'plus/more', # 'number' (only when used to replace the word number (as in telephone number) in a sentence), 2x 'twice', 100x '100 times')
- Slashes indicating that items are part of a list, an or construction or an and construction and replaced by commas, or and and respectively (e.g. locusts/grasshoppers etc. 'locusts, grasshoppers, etc.)
- However, the slash can also have other uses, namely representing two ways of saying the same thing (e.g. doctor/physician) or can represent two options that should be interchanged out in different circumstances (e.g. dog/child), in which cases the slash is kept as it is.

Questions

- Questions with no inversion (i.e. declarative structure with a question mark) are not to be normalised, but a question mark should be added to questions.
  E.g. Anyone want to talk me round and make me feel happy and warm and confident again?
- As cited above, question marks are removed from sentences that are not questions.

Euphemisms/censorship

- replace with real forms.
- E.g mf > motherfucker, f*** > fuck, frigging > ????

Emoticons and emojis

- Normalise variants of smileys for the most common:
  - :) becomes :-)
  - >:D becomes >:-D???
  - :D becomes :-D
- Emojis should be kept as they are

Sometimes things cannot be normalised

- We do not normalise idioms, expressions and dialectisms. This includes historical styles of writing (doth, thou, etc.)
- hella > no way of normalising

- We do not choose a particular variety of English for normalisation and accept the various spelling varieties that exist (i.e. British and American spellings).

Miscellaneous

- A special case: Also correct utterances if the author accidentally says the opposite of what they mean but it is understandable from the context.
  E.g. Saying the opposite of what they mean: I made Social Club and Games for Windows Live accounts but it won't work even **without (>with)** accounts

Classification of phenomena

In addition to normalisation, the different non-standard phenomena are classified. Here is the list of normalisation categories with examples.

Punctuation, typographic conventions, symbols, etc.

- **punct:diff**: extra punctuation is included or necessary punctuation is removed (e.g. missing final punctuation, missing apostrophes, commas, etc.).
    - E.g. im > I'm
- **punct:norm**: punctuation to be normalised according to certain conventions (e.g. same apostrophes and quotes).
    - E.g. that's > that's
- **caps**: capitalisation differs from what is considered standard (e.g. lowercase initial characters, all uppercase, etc.).
    - E.g. IM SO HAPPY > I'm so happy
- **slash_to_or**: a slash is used, where in normalised speech an "or" would be used to represent a list of items. This applies to the whole list, including where etc. is included. Not that this does not include cases where the items are alternatives in the discourse
    - E.g. cat/exhaust/etc > cat or exhaust, etc.
    - E.g. truth/dare > truth or dare
    - E.g. [counter-example] AW WELL MY DOG/CHILD IS VERY FRIENDLY SO LET ME APPROACH > aw, well my dog/child is very friendly, so let me approach
- **slash_to_and**: a slash is used, where in normalised speech an "and" would be used. This applies to the whole list, including where etc. is included.
    - E.g. work/paint > work and paint
- **slash_distribution**: the use of a slash to separate two items where the slash does not separate two complete items (i.e. part of one element is distributed to both items thanks to the slash). An example makes this easier to understand:
    - E.g. just disrespects any / everyone > just disrespects any / everyone
- **word_to_symbol**: the use of a symbol to represent a word
    - E.g. + > and, & > and
    - E.g. ~ > around
    - E.g. $$$ > money
- **symbol_placement**: non-standard placement of a symbol with respects to English norms.
    - E.g. 100$ > $100

Spacing

- **spacing**: missing or added spacing in the original text
    - E.g. aswell > as well
    - E.g. over thinking > overthinking
- **spacing**:**camelcase**: the use camelcase (capital letters at the beginning of words) instead of using spaces
    - E.g. surroundedUs > surrounded us
    - E.g. sawThat > saw that

Phonetically similar spellings (including imitation of speech)

- **phon**: the word uses a variant of spelling based on the phonetic similarity of the sequence of characters. This also includes the use of individual letters to represent words or syllables because of an equivalence in their pronunciation (u > you, b > be, c > see).
    - E.g. saturday **sesh** > Saturday **session** (also a case of truncation)
    - E.g. sup > What's up (also truncation)
    - E.g. bcos -> because
    - E.g. n > and
    - E.g. tho > though
    - E.g. speakin > speaking
- **phon:char**: a character is used in the place of a word or syllable because of its phonetic similarity with the word or syllable
    - E.g. b > be
    - E.g. c > see
    - E.g. u > you
- **phon:digits**: a digit is used in the place of a word or part of a word.
    - E.g. m8 > mate
    - E.g. 2 > to
    - E.g. as1 that will play > as one that will play (in a context where 1 could be incorrect, otherwise this should not be normalised)
- **phon:cute**: the spelling of a word to indicate "cute" or babyish pronunciation, e.g. using 'w' to replace an initial letter
    - E.g. wecommended > recommended
- **phon:hesitate**: words that are written in a way to imitate hesitation
    - E.g. terribl-….yyy > terribly
    - E.g. Y-y-yyy-yes > Yes
- **phon:sound**: the case of words that are used to indicate a sound (very rare)
    - E.g. bRRrrRRrrRRrr > brr rrr rrr
- **phon:interjection**: interjections that are normalised to single (and more standard) variations
    - E.g. bla > blah
    - E.g. URGHHH > ugh
    - E.g. Nawh > no

Other spelling variations (ergographic, expressiveness):

- **elongation**: characters are repeated, usually as a a mark of expressiveness.
    - E.g. *meeeeelltiiiinggg* > melting
    - E.g. sooo > so

- **devowelling**: a word with the vowels removed (initial vowels are often kept however). This can often result in double characters being reduced to single ones (messages > msgs) In this category are also words where part of the word has been devowelled.
    - E.g. wt > what, ovr > over, ppl > people
    - E.g. askd > asked (initial vowel kept)
    - E.g. travllr > traveller
- **contraction**: when several words are contracted into a single one. This has some overlap with the characteristics of phonetic distance, in that it is due to the pronunciation of the words that the contraction occurs.
    - E.g. gonna > going to
    - E.g. innit > isn't it?
- **truncation**: a word is shortened, either at the end (traditional truncation) or sometimes at the beginning, often by removing a syllable or a suffix. Note the difference with acronymisation, which involves keeping initial characters.
    - E.g. sesh > session
    - E.g. cuz > because, till > until
    - **E.g. ofc > of course -> CHANGED, NOW ACRONYM**
    - **E.g. w > with -> CHANGED, NOW ACRONYM**
- **acronym**: a word or sequence of word is represented as an acronym, i.e. the initial characters of the word (or syllables) are retained and the others are elided.
    - E.g. RN > right now
    - E.g. gf > girlfriend
    - E.g. never mind > nvm
    - E.g. w > with
- We also include in this category words that are partially acronymised (i.e. where one syllable is represented by its initial but the rest is not).is acronymised but the rest is not.
    - E.g. oline > offensive line
    - E.g. gmeet -> Google meet
    - E.g. bday > Birthday
    - E.g. ofc > of course
- Note that sometimes slashes are included in the acronym
    - E.g. b/c > because,
    - E.g. w/o > without.
- **abbreviation**: abbreviations for units of measurement and other standard cases
    - E.g. ft > feet, 2k > 2000, hrs > hours
    - E.g. Ex > for example

Spelling mistakes (distinguished from spelling variation identified as being intentional)

- **spell**: the word contains a spelling error that is not clearly intentional (covered by the other phenomena such as truncation, devowelling, etc.) and not covered by the other more specific categories.
- **spell:charswap**: the characters in the word are present but not in the right order (most often consecutive characters being swapped)
    - E.g. nobel > noble
    - E.g. furhter > futher

Misc

- **digit_letter_sim**: Very rare, but where a digit is used in the place of a letter due to the typographic similarity (see in 3ver > ever).
- **letter_to_digit**: Very rare, but where a digit is used in place of a letter not because of their typographic similarity, but because as a sort of tautology (seen in 1nce > Once).
- **suffix**: the addition of a suffix to a word, either as a diminutive or other
    - E.g. lolsky > lol
    - E.g. meanie > mean
    - E.g. doggy > dog

Added and dropped words

- **word_drop**: a word is not present in the original text and is present in the normalised version
    - E.g. It also confusing… > It's also confusing
    - E.g. u wanna see? > Do you want to see?
- **word_drop**:**pronoun**: the original text omits a pronoun (often the case of subject pronouns at the beginning of sentences) that is included in the normalised version.
    - E.g. Was gunna try distortion… > I was going to try distortion…
- **word_drop**:**det**: the original text omits an article (e.g. the or a) that is included in the normalised version.
    - E.g. Pretty creative way… > A pretty create way…
- **word_add**: a word is present in the original text and is removed in the normalised version
    - E.g. … in ten days ago > …ten days ago
    - E.g. also for uses of word "like" as a filler
- **word_add**:**det**: the original text includes an article where the normalised version removes it
    - E.g. …adds an 12kg of salt > …adds 12kg of salt
- **symbol_drop**: the original text omits a symbol that is included in the normalised version.
    - E.g. 32c > 32°C
- **symbol_add**: the original text includes a symbol that is removed in the normalised version.
    - …no issue w being over 12+ ft… > …no issue with being over 12 feet…

Grammar

- **inflection**: a word is not correctly inflected (e.g. with respect to number, tense, etc.)
    - E.g. …wondering what ppl **thought** are > …wondering what people's **thoughts** are
- **grammar**: inflection-related errors
    - E.g. …wondering what **ppl** thought are > …wondering what **people's** thoughts are
    - E.g. if **your** good > if **you're** good
- **grammar:v**:
- **grammar:v:inflect**

Lexical changes

- **lex_choice**: a use of a non-standard lexical choice, including dialectisms (e.g. cannae, ain't), malapropisms (e.g. genually), foreign words and generally wrong choices of words (e.g. wrong part of speech, wrong semantic choice of words, lacking punctuation, use of an antonym by accident, etc.)
    - E.g. I am confusion > I am confused
    - E.g. genually > genuinely
    - E.g. pish > piss
    - E.g. ain't > aren't

    - E.g. cannae > cannot
    - E.g. y'all > everyone/all/all your (depending on the context)
    - E.g. sans guac > without guacamole
- **surrounding_emphasis**: emphasis added to certain words typographically (removed in the normalised variants).
    - E.g. *without* > without
    - E.g. ~find~ > find
- **emoticon**: emoticon that is a variant on the common emoticons :-), :-D, :-(, :-/ and >:-)
    - E.g. :-///// > :-/
    - E.g. (: > :-)
    - E.g; :^) > :-)
- **censored**: the word contains symbols in an effort to censor the word
    - E.g. upv*te > upvote
    - E.g. s**t > shit, sh** > shit

Changes made since v1

Made distinction between categories clearer (acronyms, truncation, contraction) and cleaning up of annotations.

Renaming of categories for brevity and clarity purposes
- phonetic_distance > phon
- spelling_error > spell
- digits_to_words > digit
- scrambled > charswap
- mimic_spoken > hesitate
- acronymisation > acronym
- capitalisation > caps
- censure > censored

Some categories restructured so that they are part of larger categories (for coherence)
- digit > phon:digit
- cute > phon:cute (wecommended > recommended) and also suffix (doggy > dog)
- camelcase > spacing:camelcase
- charswap > spell:charswap
- hesitate > phon:hesitate
- placeholder > diff_punct
- article_drop > word_drop:det
- pronoun_drop > word_drop:pronoun
- article_add > word_add:det
- norm_punct > punct:norm
- punct_diff: punct:diff
- inflection > grammar and more fine-grained categories introduced used grammar
- dialectism > lex_choice
- foreign > lex_choice (given that no truly code-switched text is included in the test set)
- interjection > phon:interjection

• grammar and inflection > grammar (with subcategories added)

Removal of categories

• double_to_single_character (all cases occur with devowelling so could be seen as part of the devowelling process)

Spans of phenomena

Added words

- Span of the addition concerning spaces:
  The addition of a word often (although not always) has the effect of adding an additional space added to the sentence. The question of whether the "added space" is the one that is found to the right or left of the word is not theoretically pertinent, because spaces are simply typographic conventions (i.e. the identity of the original space cannot be traced to either the right or left space after the addition of the extra word). However, it is necessary to decide on a annotation convention when annotating the addition of the word and the accompanying space. It is not possible to say that the "added space" is systematically considered to be the one to the right/left, because words are not always surrounded by spaces (e.g. at the beginning or at the end of the sentence, a space is necessarily added to the right or the left respectively, and a similar case is following or preceding punctuation). We therefore consider the "added space" to be the one that "links" the added word most appropriately to its syntactic constituent. For example, the span for an added pronoun would include the space to the right, thus linking it to the following verb (idem for an added article). However, an added preposition making up a preposition verb would include the space to the left, thus linking it to the preceding verb.
- Effect on capitalisation:
  An added word at the beginning of a sentence (e.g. a pronoun in a pro-drop scenario), can affect what is expected of capitalisation from the word that follows (i.e. previously the initial word of the sentence). For example, if the raw sentence starts with "Said he would…" and the normalised version is "He said he would", then there has been an addition of an initial pronoun. The word "said" has undergone a change in capitalisation as a result of the addition of a pronoun, due to the shifting of the capital letter to the added pronoun.