Creating Data in Icelandic for Text Normalization

Helga Svala Sigurðardóttir
Reykjavik University
helgas@ru.is

Anna Björk Nikulásdóttir
Grammatek
anna@grammatek.com

Jón Guðnason
Reykjavik University
jg@ru.is

Abstract

We introduce Regína, a rule-based system that can automatically normalize data for a text-to-speech (TTS) system. Normalized data do not generally exist so we created good enough data for more advanced methods in text normalization (TN). We manually annotated the first normalized corpus in Icelandic, 40,000 sentences, and developed Regína, a TN-system based on regular expressions. The new system gets 89.82% accuracy compared to the manually annotated corpus on non-standard words and showed a significant improvement in accuracy when compared to an older normalization system for Icelandic. The normalized corpus and Regína will be released as open source.

1 Introduction

Text normalization is an integral part of a TTS system. Unrestricted input texts can contain so-called non-standard words (NSWs), which are impossible for a computer to read without being formatted into regular strings of alphabetical letters and punctuation marks. These NSWs are divided into semiotic classes and include abbreviations, numbers, and special characters.

The degree of importance of text normalization in TTS is not obvious even though its utility is known. Most words do not need to be normalized, and therefore normalized datasets and their unnormalized counterparts are almost identical. However, without expanding NSWs, a TTS system skips those words, making the text inaccurate and incomplete.

To clarify, let us look at an example of a sentence before and after normalization.

```
 Hæsti tindur Esjunnar er 914 m.
(Esjan's highest peak is 914m.)
```

↓

```
 Hæsti tindur Esjunnar er níu
hundruð og fjórtán metrar.
(Esjan's highest peak is nine
hundred and fourteen meters.)
```

Text normalization systems are customarily rule-based but are moving in the direction of neural networks (NNs). Models made with NNs require less human effort (Graves and Jaitly, 2014) but need a vast amount of correctly annotated data to learn from, and these do not naturally exist for text normalization. People can generally read NSWs without requiring an explanation, so there is no motivation to create data with normalized text, such as in translation. To acquire data in Icelandic for the training of more sophisticated systems, we start by making a system that can make data good enough for further training. We compare the results of this system with manually annotated data to better assess the quality.

1.1 Background

In 1996, Sproat (Sproat, 1996) published work for a unifying model for most text normalization problems, built with Weighted Finite-State Transducers (WFSTs). The transducers were constructed using a lexical toolkit that allows descriptions of lexicons, morphological rules, numeral-expansion rules, and phonological rules. In 2001, Sproat (Sproat et al., 2001) expanded on this work and described challenges that heavily inflected languages like Russian (and Icelandic) face. This work was the first that treated the problem as essentially a language modelling problem.

Up until recently, the primary approach to the text normalization problem was with WFSTs. In 2015, Ebden et al. (Ebden and Sproat, 2015) released a paper where they described the Kestrel text normalization system, a component of the Google TTS system. It differed from previous systems by separating the tokenization and classification (determining whether a word should be normalized and, if so, which semiotic class it belongs to) from the verbalization step. Kestrel recognizes a large set of semiotic classes: various categories of numbers, times, telephone numbers and electronic addresses.

Work on Icelandic spoken language technologies is defined within the Language Technology Programme for Icelandic (2019-2023) (Nikulásdóttir et al., 2020). Previous work on language resources for Automatic Speech Recognition (ASR) and TTS include acoustic data gathering (Guðnason et al., 2012; Steingrímsson et al., 2017; Mollberg et al., 2020) and text corpus building for Icelandic (Steingrímsson et al., 2018). Spoken language technologies for Icelandic commenced with building ASR systems (Helgadóttir et al., 2017) with resource work on TTS aimed at a pronunciation lexicon (Nikulásdóttir et al., 2018) and acoustic data recordings (Sigurgeirsson et al., 2020) following.

The only research that has been done on text normalization in Icelandic was done in 2019, (Nikulásdóttir and Guðnason, 2019) focusing exclusively on numbers. The system built follows the open-source version of Kestrel, Sparrowhawk[1] (Ebden and Sproat, 2015), and contains a set of grammar rules written in Thrax. Numbers are handled with a classification grammar, which classifies input containing digits into several semiotic classes, and a verbalization grammar, which inflates the numbers. The verbalization grammar labels possible verbalizations with part-of-speech tags and a language model is then used to choose the most probable word form where verbalization is ambiguous.

In the last few years, people have been experimenting with deep learning (neural networks) for text normalization (Pusateri et al., 2017; Pramanik and Hussain, 2019; Zhang et al., 2019). This works well for many tasks, but the task of text normalization is fragile. Neural networks are prone to so-called unrecoverable errors; they do not only expand the words incorrectly, but the result is misleading. For instance, a navigation system could send the user to another side of town because it incorrectly expanded the postal code. Some experiments have been performed with hybrid systems, using a neural model and then applying a grammar system, such as Kestrel. The grammar system implements an overgenerating grammar, which includes the correct verbalization, and can be used to guide the system (Sproat and Jaitly, 2017; Zhang et al., 2019, 2020).

In 2016, Sproat et al. (Sproat and Jaitly, 2016) released a challenge: given a large corpus of written text aligned to its normalized spoken form, train an RNN to learn the correct normalization function. The authors presented a dataset of general text with generated normalizations using an existing text normalization component of a TTS system (Kestrel).

2 Data

The data used are 40,000 sentences (741,909 words) from the 2017 version of the Icelandic Gigaword Corpus (IGC). We use sentences that include many NSWs, such as numbers, abbreviations, and symbols. They are from all sources in the IGC. 534 of the sentences deal with sports results and were handled separately. The sentences were manually annotated and make up the first manually curated normalization corpus for Icelandic. For a small experiment on inter annotator agreement, three people from Reykjavík University normalized 30 sentences with 205 NSWs, using the guidelines in Appendix B. The annotators expanded words without regard to a semiotic class. The inter-annotator agreement for NSWs was $\kappa = 0.85$.

3 Methodology

Icelandic is an inflected language, where each word can have various forms of words depending on the context. For example, the number *2 (two)* can be expanded as *tveir, tvo, tveimur, tveggja, tvær*, or *tvö*, depending on the next word's case. The ordinal number *2. (second)* can then be *annar, annan, öðrum, annars, önnur, aðra, annarri, annarrar, annað, öðru, annars, aðrir, annarra*, or *aðrar*. Only the first four numbers (one, two, three, and four) have this inflected nature.

The most significant ambiguity in the data was whether to write hyphens and dashes as *til* (to) or silence when it was used to describe sports results. In Icelandic, a sentence like *Leiknum lauk með 2-1 sigri* (The game ended with a 2-1 victory), is read as *Leiknum lauk með tvö (2) eitt (1) sigri* and the hyphen is silent. In a TTS system, the idea is that the user can either mark the topic herself or run the text through data-driven topic classification.

[1] https://github.com/google/sparrowhawk

The system built in this research uses regular expressions and grammar rules to determine how a word should be expanded. It has been given the name Regína. The first step of Regína is to run rules for expansions of abbreviations, measurements, money, weblinks, and roman numerals through the unnormalized text. The rules for measurements take prepositions into account. For example, this could help when the base version of *km* is *kílómetrar*. If we say *til 2 km*, Regína uses the preposition *til* to expand the word to the genitive case, *kílómetra*. The next step is to run this expanded text through a part-of-speech (POS) tagger. (Steingrímsson et al., 2019) Instead of reading *km* as an abbreviation (and giving it a tag as such), the tagger now recognizes the word *kílómetra* and knows from context it is in genitive case. Now Regína is preserving part-of-speech tags for each word. Next, the semiotic class of remaining NSWs is determined. Rules for numbers are applied to cardinal and ordinal numbers, decimals and fractions. In this step, the words tagged as numbers consider the next word's tag. The numbers that are not followed by an adjective or a noun are assigned a default case. The final step of the system is to run the text through rules for other semiotic classes: time, sports results, digits, letters, dates, and symbols. For comparison, the normalized text was re-aligned with the manually annotated text, with each sentence and word indexed to keep the structure clear. In Appendix A, the pipeline for Regína is shown.

4 Results

The dataset with general news had 729,763 words, of which 701,088 did not need normalization. The baseline of the system without any work was thus 96.08%. The remaining 28,675 words were split into cardinal, ordinal, and decimal numbers, digits, fractions, letter sequences, abbreviations, weblinks, measurements, clock times, dates, and symbols. The accuracy and size of each class are shown in Table 1.

Sports

The only specific domain looked at were sports because of the ambiguity regarding hyphens. The portion regarding sports was 12,106 words, 1.7% of the dataset. The ratio of NSWs in need of normalization is relatively high in sports, 14.66%. We looked at the same semiotic classes, with an addition of a special one for sports results.

SEMIOTIC CLASS	**ACCURACY [%]**	**# examples**
ALL	99.51	729,673
PLAIN	99.94	626,541
CARDINAL	86.87	8,456
ORDINAL	87.24	1,653
DIGIT	51.45	241
DECIMAL	74.36	197
FRACTION	33.33	39
LETTERS	96.05	3,576
ABBREVIATIONS	80.72	1,675
ROMAN NUMERALS	33.66	104
MONEY	46.89	352
WLINK	99.14	348
MEASURE	61.96	1,559
TIME	80.36	713
DATE	97.75	7,937
SYMB	88.36	1,735
PUNCT	99.93	74,547

Table 1: Results for general news

SEMIOTIC CLASS	**ACCURACY [%]**	**# examples**
ALL	98.45	12,106
PLAIN	99.98	8,923
CARDINAL	96.84	538
ORDINAL	91.89	74
DIGIT	0.0	1
DECIMAL	0.67	3
FRACTION	0.0	1
LETTERS	99.06	106
ABBREVIATIONS	75.0	20
WLINK	100.0	1
MEASURE	60.0	5
TIME	100.0	2
DATE	88.4	43
SYMB	90.91	88
SPORT	84.55	893
PUNCT	1.0	1,408

Table 2: Results for sports news

Error division

We considered error division for the classes and listed them in Table 4. All classes are handled alike in the two domains except for the symbol class (where a dash is generally a *til (to)* but silent in the sport domain), and the SPORT class is unique to sports news. The errors are divided up to:

- *CLASS* – incorrect normalization due to misclassification of the token
- *FORM* – incorrect grammatical form of the normalization but otherwise correct
- *NON-ERRORS* – errors due to errors in the manual data, misalignment of whitespaces, or instances where both expansions are correct but different (e.g. þúsund and eitt þúsund (thousand and one thousand)).

SEMIOTIC CLASS	ORIGINAL	MANUAL	MACHINE (evt. classification)	ERROR
CARDINAL	4	fjórum	fjögur	FORM
ORDINAL	2.	öðru	annað	FORM
DECIMAL	2.4	tveir komma fjórir	tveir punktur fjórir (DIGIT)	CLASS
DECIMAL	12,883	tólf þúsund átta hundruð áttatíu og þrír	einn fimm komma átta átta þrír (DIGIT)	CLASS
DATE	4/4	fjórði apríl	fjórir fjórðu (FRACTION)	CLASS
FRACTION	1/8	einum áttunda	einn áttundu	FORM
PLAIN	ALLIR	ALLIR	A L L I R (LETTERS)	CLASS
ABBREVIATION	-100 kg	undir hundrað kíló	mínus hundrað kíló (wrong word)	OTHER
CARDINAL	70s	seventies (English)	sjötíu sekúndur	OTHER
MEASURE	3 cm	þriggja sentimetra	þrír sentimetrar	FORM
TIME	1:22	eitt tuttugu og tvö	ein tuttugu og tvær	FORM
DATE	1. nóv 2012	fyrsta nóvember tvö þúsund og tólf	fyrsti nóvember tvö þúsund og tólf	FORM
SPORT	24/7	tuttugu og fjóra <sil> sjö	tuttugu og fjögur <sil> sjö	FORM
SYMB (general)	-	-	til	OTHER
SYMB (sport)	-	til	-	OTHER
PUNCT	/	/	skástrik (SYMB)	CLASS

Table 3: Incorrect results from Regína

- *NO ACTION* – the token was not expanded
- *INSUFFICIENT* – the token was only partially expanded
- *OTHER* – the token was normalized incorrectly, not due to class or grammatical form. Examples include dates written in English, incorrectly expanded dashes, and reverse order of money, such as $5 incorrectly being expanded to *dollarar fimm (dollars five).*

Comparison with an existing system

To compare Regína with the old Thrax normalizer, Textahaukur (Nikulásdóttir and Guðnason, 2019), 400 sentences from the whole dataset were normalized with both systems. 147 of those contained NSWs and were observed for more meaningful results. Regína had an accuracy score of 83.67%, with 20 sentences containing 22 words that did not match the manual annotation. Textahaukur had an accuracy score of 61.22%, with 55 sentences containing 106 incorrectly normalized words.

4.1 Discussion

Normalization systems are either rule-based, made with neural models or a hybrid of those two. The drawback of a rule-based system is that it is less generalizable and requires more maintenance. The main advantage is that it never makes *unrecoverable errors*. The worst errors Regína makes is not expanding a non-standard word, which happens when it does not find an appropriate semiotic class. It can also happen that it assigns the wrong class to it – making the expansion comprehensible but awkward.

As mentioned, the main problem with an inflected language like Icelandic is that each word has several forms. A part-of-speech tagger helps determine the expansion of the preceding number, but if the word following a number is not a noun or an adjective, it is given a default form. For cardinal numbers, that is the neutral, nominative, singular version, which works well with sports results, years, timings, addresses, et cetera. For decimals, it is the masculine, nominative, singular version. For ordinal numbers, it is the masculine, dative, singular form. This covers most cases, especially dates.

These default cases, plus the next word's tag, covered a vast majority of examples in the data. The incorrect examples from these semiotic classes, as seen in 4, are mostly from the target word neither having a tag for reference nor being in the default form. Abbreviations, measurements, and fractions have the same problem, i.e., the default class is not correct. The system also marks dates written as *6/6* as fractions and expands them to *sex sjöttu (six sixths)* instead of *sjötti júní (the sixth of June)*.

The system is built with an intention of a spell-correcting layer before the normalization. In Icelandic, the rule is to write thousands separators with a dot and decimal separators with a comma, opposite to English. Regína sends numbers that do not conform to Icelandic rules to the digit class and writes them out, digit by digit, sometimes going against the author's intention.

The time class only has rules for the 24-hour clock format, so when it read results from time-keeping, it did not expand the numbers correctly. The symbol class mostly suffers from the strict

SEMIOTIC CLASS	# ERRORS	CLASS	FORM	NON-ERRORS	NO ACTION	INSUFFICIENT	OTHER
PLAIN	384	280	0	103	0	0	1
CARDINAL	882	17	820	23	8	13	1
ORDINAL	223	6	212	0	0	5	0
DIGIT	118	110	0	4	0	4	0
DECIMAL	51	27	23	0	1	0	0
FRACTION	27	3	21	0	1	0	2
LETTERS	142	9	0	11	120	2	0
ABBREVIATIONS	328	51	83	1	184	8	1
ROMAN NUMBERS	69	0	47	0	22	0	2
MONEY	188	1	84	3	5	64	31
WLINK	4	2	0	0	1	0	1
MEASURE	595	6	524	2	0	60	3
TIME	140	4	3	1	0	132	0
DATE	182	16	81	663	4	8	11
SPORT	140	46	58	34	2	0	0
SYMB (general)	202	0	1	1	189	0	11
SYMB (sport)	8	0	0	0	0	0	8
PUNCT	53	50	0	3	0	0	0

Table 4: Error division

translation of / to *skástrik* (slash) and *-/–* to *til* in general text, silence in sports. Regína tried to catch all non-standard words, sometimes outside its scope. Parts of sentences in Icelandic text are sometimes written with spelling errors, in English, or as with the separators, with rules that do not apply to Icelandic. Both ends have rigid rules about weblinks and sports results, and the results are almost 100% accurate. The only incorrect examples are misclassified – like 24/7 (twenty-four-seven) is classified as a sports result.

Finally, the slight inaccuracy of the plain class, which should remain unchanged, resulted mainly from words being misclassified to the LETTERS class (NATO → N A T O) and mistakes in the manual data.

4.1.1 Comparison between systems

The errors made by Regína and Textahaukur were examined. Regína had some abbreviations that were not expanded because of possible ambiguity. Otherwise, a majority of the errors was the wrong case of an expanded number.

These were also the most common errors for Textahaukur. More serious errors were a strong tendency to change cases in the middle of a token. For example, the number 110 was normalized in the feminine for the first part (hundraðasta og) and then masculine (tíundi). Textahaukur deleted tokens when they were followed by a token it could not handle (5,5°C became °) or skipped handling a whole sentence. In some cases, Textahaukur did not have any rules implemented. These were cases of weblinks and sports, which Regína handles almost perfectly with rigid rules on both ends. Regína and Textahaukur both had cases where they expanded correctly, but the manual normalization was incorrect, showing that even when a computer knows less than a person, it is more consistent.

4.2 Conclusions and future work

Regína works well and does not return misleading results. The manually annotated data inevitably became a development dataset, since it was always visible for the developer of Regína. However, this is exclusively a problem for comparing the system with the corpus. Regína will be used to normalize text for TTS synthesis. Although the exact expansion might differ from person to person, that does not indicate an incorrect normalization.

In the future, we want to do more thorough experiments on inter-annotator agreement. For the 205 words, the annotators mostly disagreed on words that can be expanded in multiple ways. Regína will be used to normalize more data for further development in text normalization, using neural models. For the TTS application, we will create a test set-up for extrinsic evaluation given the new dataset.

Acknowledgements

This work is supported by the Language Technology Programme for Icelandic 2019-2023, funded by the Icelandic government. We want to thank Haukur Páll Jónsson and David Erik Mollberg for annotation and conversations about data structure, and Ari Páll Kristinsson, for advice on grammar.

References

Peter Ebden and Richard Sproat. 2015. The Kestrel TTS text normalization system. *Natural Language Engineering*, 21(3):333.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

Jón Guðnason, Oddur Kjartansson, Jökull Jóhannsson, Elín Carstensdóttir, Hannes Högni Vilhjálmsson, Hrafn Loftsson, Sigrún Helgadóttir, Kristín M Jóhannsdóttir, and Eiríkur Rögnvaldsson. 2012. Almannaromur: An open icelandic speech corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2017. Building an asr corpus using althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jon Gudnason. 2020. Samrómur: Crowd-sourcing data collection for icelandic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3463–3467.

Anna Björk Nikulásdóttir and Jón Guðnason. 2019. Bootstrapping a Text Normalization System for an Inflected Language. Numbers as a Test Case. In *INTERSPEECH*, pages 4455–4459.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for icelandic 2019-2023. *arXiv preprint arXiv:2003.09244*.

Anna Björk Nikulásdóttir, Jón Guðnason, and Eiríkur Rögnvaldsson. 2018. An icelandic pronunciation dictionary for tts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.

Subhojeet Pramanik and Aman Hussain. 2019. Text normalization using memory augmented neural networks. *Speech Communication*, 109:15–23.

Ernest Pusateri, Bharat Ram Ambati, Elizabeth Brooks, Ondrej Platek, Donald McAllaster, and Venki Nagesha. 2017. A Mostly Data-Driven Approach to Inverse Text Normalizatio7n. In *INTERSPEECH*, pages 2784–2788. Stockholm.

Atli Sigurgeirsson, Gunnar Örnólfsson, and Jón Guðnason. 2020. Manual speech synthesis data acquisition-from script design to recording speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 316–320.

Richard Sproat. 1996. Multilingual Text Analysis for Text-to-Speech Synthesis. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1365–1368. IEEE.

Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer speech & language*, 15(3):287–333.

Richard Sproat and Navdeep Jaitly. 2016. RNN Approaches to Text Normalization: A Challenge. *arXiv preprint arXiv:1611.00068*.

Richard Sproat and Navdeep Jaitly. 2017. An RNN Model of Text Normalization. In *INTERSPEECH*, pages 754–758. Stockholm.

Steinþór Steingrímsson, Jón Guðnason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2017. Málrómur: A manually verified corpus of recorded icelandic speech. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 237–240.

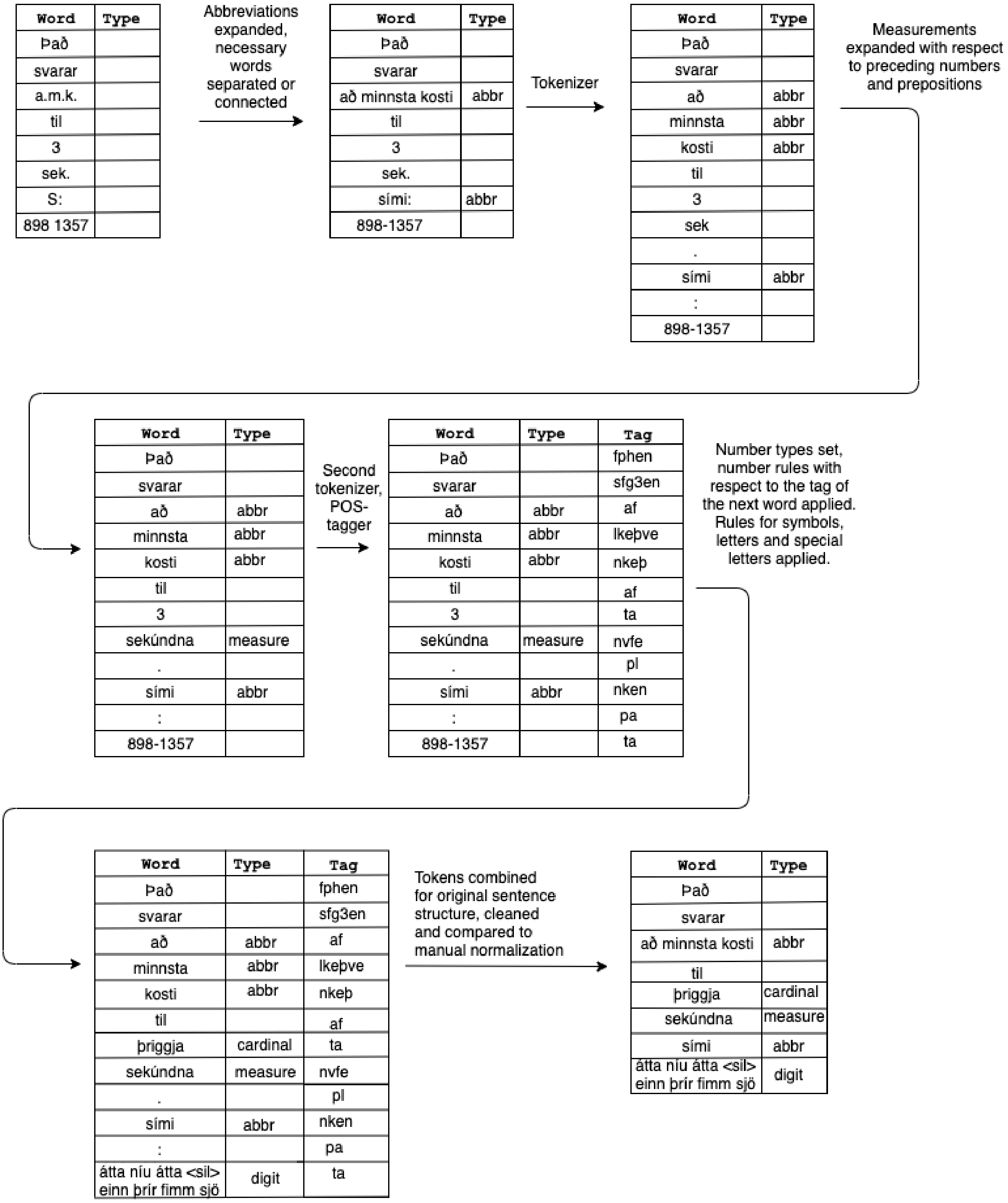
Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a bilstm tagger with a morphological lexicon and a lexical category identification step. *arXiv preprint arXiv:1907.09038*.

Hao Zhang, Richard Sproat, Axel H Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural Models of Text Normalization for Speech Applications. *Computational Linguistics*, 45(2):293–337.

Junhui Zhang, Junjie Pan, Xiang Yin, Chen Li, Shichao Liu, Yang Zhang, Yuxuan Wang, and Zejun Ma. 2020. A hybrid text normalization system using multi-head self-attention for mandarin. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6694–6698. IEEE.

A Regína Pipeline



Word	Type
Það	
svarar	
a.m.k.	
til	
3	
sek.	
S:	
898 1357	

Abbreviations expanded, necessary words separated or connected

Word	Type
Það	
svarar	
að minnsta kosti	abbr
til	
3	
sek.	
sími:	abbr
898-1357	

Tokenizer

Word	Type
Það	
svarar	
að	abbr
minnsta	abbr
kosti	abbr
til	
3	
sek	
.	
sími	abbr
:	
898-1357	

Measurements expanded with respect to preceding numbers and prepositions

Word	Type
Það	
svarar	
að	abbr
minnsta	abbr
kosti	abbr
til	
3	
sekúndna	measure
.	
sími	abbr
:	
898-1357	

Second tokenizer, POS-tagger

Word	Type	Tag
Það		fphen
svarar		sfg3en
að	abbr	af
minnsta	abbr	lkeþve
kosti	abbr	nkeþ
til		af
3		ta
sekúndna	measure	nvfe
.		pl
sími	abbr	nken
:		pa
898-1357		ta

Number types set, number rules with respect to the tag of the next word applied. Rules for symbols, letters and special letters applied.

Word	Type	Tag
Það		fphen
svarar		sfg3en
að	abbr	af
minnsta	abbr	lkeþve
kosti	abbr	nkeþ
til		af
þriggja	cardinal	ta
sekúndna	measure	nvfe
.		pl
sími	abbr	nken
:		pa
átta níu átta <sil> einn þrír fimm sjö	digit	ta

Tokens combined for original sentence structure, cleaned and compared to manual normalization

Word	Type
Það	
svarar	
að minnsta kosti	abbr
til	
þriggja	cardinal
sekúndna	measure
sími	abbr
átta níu átta <sil> einn þrír fimm sjö	digit

Figure 1: Pipeline of Regína from unnormalized to normalized text.

B Normalization guidelines

SEM. CLASS	EXPLANATION	EXAMPLE	NORMALIZED
PLAIN	Words remain same	dæmum	dæmum
PUNCT	Punctuation marks	.,?!:;”„“...·|	.,?!:;”„“...·|
CARDINAL	Cardinal numbers	86.761	áttatíu og sex þúsund sjö hundruð sextíu og einn/ein/eitt/eina/einum /einni/einu/eins/einnar
		337.429	þrjú hundruð þrjátíu og sjö þúsund fjögur hundruð tuttugu og níu
ORDINAL	Ordinal numbers	86.761.	áttatíu og sex þúsund sjö hundruð sextugasti og fyrsti / sextugasta og fyrsta/sextugustu og fyrstu
		337.429.	þrjú hundruð þrjátíu og sjö þúsund fjögur hundruð tuttugasti og níundi /tuttugasta og níunda/tuttugustu og níundu
LETTERS	Letter sequences	KR	K R (ká err)
		ehf	E H F (e há eff)
DATE	Dates	1919	nítján hundruð og nítján
		29. september 1928	tuttugasti/a og níundi/a september nítján hundruð tuttugu og átta
		14. mars	fjórtándi/a mars
		september 2008	september tvö þúsund og átta
		kl. 20:00	klukkan tuttugu núll núll
		klukkan 11.15	klukkan ellefu fimmtán
MEASURE	Measurements	120 kW	hundrað og tuttugu kíló(vött/vöttum/vatta)
		5%	fimm prósent(um/a)
		39,5 kg	þrjátíu og níu komma fimm kíló(um/a)
SYMB	Symbols	+	plús
		-	mínus
		@	hjá
		©	höfundarréttur
ABBR	Abbreviations	a.m.k.	að minnsta kosti
		SV-átt	suðvestanátt
WLINK	Web handles	helgas@ru.is	h e l g a s hjá r u punktur i s
		@BarackObama	hjá B A R A C K O B A M A
		#ljosanott2014	myllumerki l j o s a n o t t tveir núll einn fjórir
DECIMAL	Decimal numbers	0,45	núll komma fjórir/fjóra/fjórum /fjögurra/fjórar/fjögur fimm

SEM. CLASS	EXPLANATION	EXAMPLE	NORMALIZED
SPORT	Sports results	2-1 3:0 16/5 (fráköst)	tvö eitt þrjú núll sextán <sil> fimm (fráköst)
RNUM	Roman numerals	XII	tólf(ti/ta/tu)
FRACTION	Fractions	½ 2/6 1 1/3	hálfur/hálfan/hálfum/hálfs/hálf/hálfa /hálfri/hálfrar/hálft/hálft/hálfu tveir/tvo/tveimur/tveggja/tvær/tvö sjöttu einn og einn þriðji / einn og einn þriðja / einum og einum þriðja/eins og eins þriðja / ein og ein þriðja / eina og eina þriðju / einni og einni þriðju / einnar og einnar þriðju / eitt og eitt þriðja/einu og einu þriðja / eins og eins þriðja
DIGIT	Digit numbers	1109-05-420 365	einn einn núll níu <sil> núll fimm <sil> fjórir tveir núll þrír sex fimm
MONEY	Monetary amounts	3000 kr. kr. 4000 $40 38 m.kr.	þrjú þúsund krónur/krónum/króna fjögur þúsund krónur/krónum/króna fjörutíu dollara(r/um) þrjátíu og átta milljón(ir/um/a) króna

B.1 Rules

- Separate a word that's built from letters and numbers, C19 becomes C nítján, 1.ferð becomes 1. ferð → fyrsta ferð.
- Delete a dash at the start of the line.
- If a word ends in dash it is ignored.
- @ is written *hjá*.
- = is written *jafnt og*.
- Links are written like *www.mbl.is/123 → w w w punktur m b l punktur i s* skástrik einn tveir þrír, all letters are separated except for symbols and numbers, they are written out.
- For basketball results like *24/14 fráköst*, the */* is written as <*sil*>, i.e., *24/14 fráköst → tuttugu og fjögur <sil> fjórtán fráköst*.
- In digit sequences, dashes are written as <*sil*>, e.g., *234-353-42 → tveir þrír fjórir <sil> þrír fimm þrír <sil> fjórir tveir*

B.2 Ambiguities

- **DASH**: can imply *bandstrik (dash)* (links), <*sil*> (sports results), *til* (number intervals) or nothing.
- **SLASH**: can imply *skástrik (slash)* (links), *og (and*, *eða (or)*, a fraction, a <*sil*> or nothing.