Russian Diminutive Nouns with Logistic Regression

Elena Badmaeva
s2805820

Index

- Introduction
- Diminutive suffixes for nouns
- Model
- TiMBL
- MGL
- Logistic Regression
- Conclusion
- References

Introduction

Bauer et al.: “The notion of diminutive [...] is not easy to define clearly. One problem with this notion is the semantics, the other the kind of formal means employed to express diminutive meaning.”

Diminuitve suffixes for nouns:

Maculine gender:

-ик, -ок, -ёк (-ik, -ok, -yok)

нос (nos) - носик (nosik) ‘nose’

Feminine and neutral gender:

-чк-, -шк-, -оньк- or **-еньк-** (-chk-, -shk-, -on'k-, -en'k-)

вода (voda) - водичка (vodichka) ‘water’

солнце (solntse) - солнышко (solnishko) ‘sun’

изба (izba) - избушка (izbushka) ‘traditional russian wooden house’

Difficulties

Various diminutive forms for one word:

мама (mama, mom):

- **мамочка**(mamochka, affectionate sense)
- **мамуля** (mamulya, affectionate and playful sense)
- **маменька** (mamen'ka, affectionate and old-fashioned)
- **маманя** (mamanya, affectionate but disdainful)

Difficulties

Combination of several diminutive suffixes to make several degrees of diminution:

- **пирог** (pirog, a pie) becomes **пирожок** (pirozhok, a small pie, or an affectionate name), which then may become **пирожочек** (pirozhochek, a very small pie, or an affectionate name)

- **сыр** (syr, cheese), **сырок** (syrok, an affectionate name or a name of a small packed piece of cheese), **сырочек** (syrochek, an affectionate name)

Difficulties

Often formative infixes and suffixes look like diminutive:

- **сота** (sota, a honeycomb) and **сотка** (sotka, one hundred sqr. meter)
- **труба** (truba, a tube) and **трубка** (trubka, a special kind of a tube: telephone receiver, TV tube, tobacco pipe - in all these cases there is no diminutive sense). However, **трубка** also means a small tube (depending on context)

The word **конь** (kon', a male horse) has a diminutive form **конёк** (koniok). But **конёк** (koniok) also means a skate (ice-skating, no diminutive sense in this case), and has another diminutive form **конёчек** (koniochek, a small skate). The word **конёк** also means a gable with no diminutive sense.

Model

Input suffixes:

-Ce, Cek, Cik, Cka, eCka, eNka, eZki, iCe, iCek, ik, iSko, ka, oCka, ok, onka, Ta, uSka, yonka, Zenka, Zka, Zok.

TiMBL

Memory-Based Learning (MBL) is an elegantly simple and robust machine-learning method applicable to a wide range of tasks in Natural Language Processing. It is based on the idea that learning and processing are two sides of the same coin (Daelemans 2009). Learning is the storage of examples in memory, and processing is similarly based reasoning with these stored examples. The advantage of TiMBL is that it can automatically detect the importance of each variable in the classification of the training set.

Features for TiMBL

- phonological structure of the last three syllables;
- stress;
- last sound.

Results of TiMBL

TiMBL - Memory-Based Learning

Overall accuracy: 0.666667 (38/57)

There were 5 ties of which 2 (40.00%) were correctly resolved

Conclusion

- one-syllable words are correctly predicted
- masculine gender is easier to predict
- confusion of words with two consonants in the end

MGL

Minimal Generalization Learner or MGL is an analogical model that derives rules from an input set of words (Albright and Hayes 2002). The input of the MGL model consists of form pairs: a base form and a derived form, for example, in this project, such as regular noun and diminutive form.

The Minimal Generalization Learner constructs a large set of weighted rules that are learned during training (Goldsmith 2011). Once learning is completed, the rules are applied online during testing. This way the program ends up for each outcome with a system of interrelated rules ranging from the very specific to the very general.

Results of MGL

In total 251 nouns were used for training and 50 for testing the model. MGL model achieved much higher accuracy than TiMBL, only 6 words were predicted incorrectly.

Logistic Regression

Logistic regression is a type of regression used when the dependant variable is binary or ordinal (e.g. here is the outcome is either “correct” or “incorrect”).

- The dependent variable (prediction) is 0 or 1:

  Did the model predict the diminutive form correctly?

- The logistic regression equation is solved iteratively (by R):

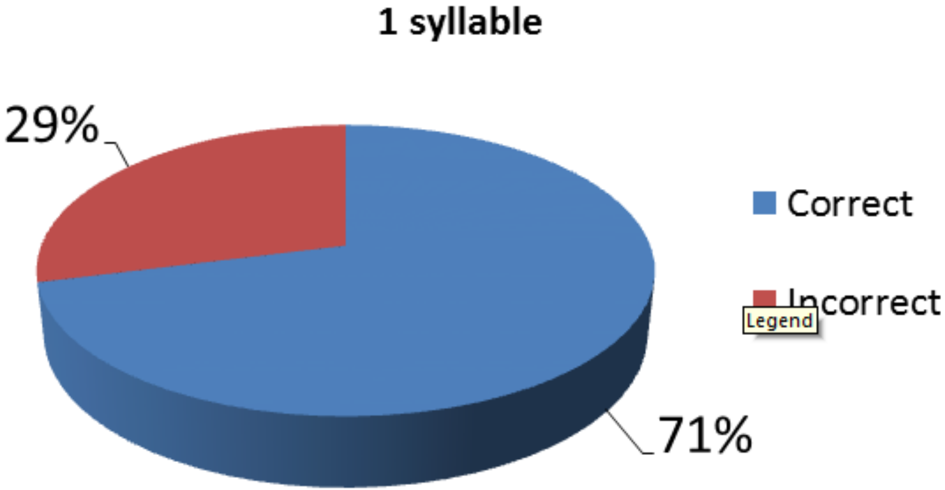
  A trial equation is fitted and tweaked over and over to improve the fit; iterations stop when improvement stops (here 16 iterations were executed)

Sample of Input Data

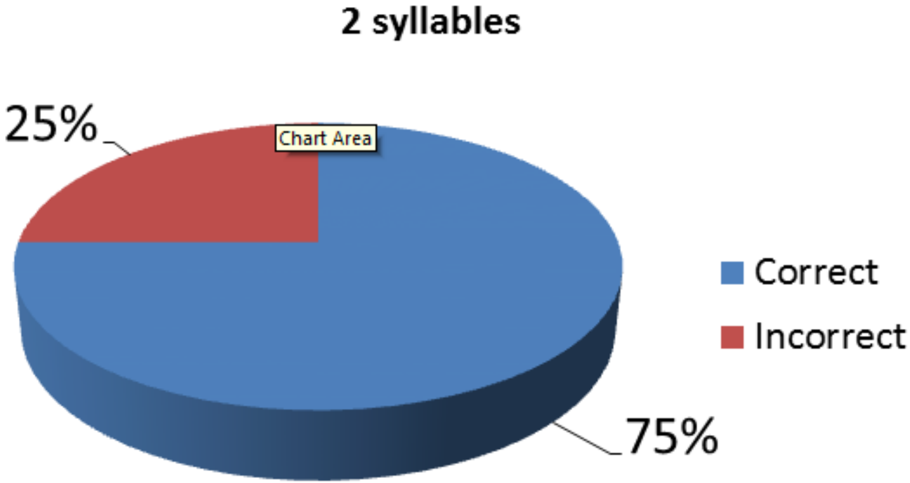
```
word,correctness,number_of_syllables,last_sound
bereg,1,2,g
brat,1,1,t
bumaga,1,3,a
Cas,1,1,a
Celovek,0,3,k
DaDa,1,2,a
delo,1,2,o
```

Results

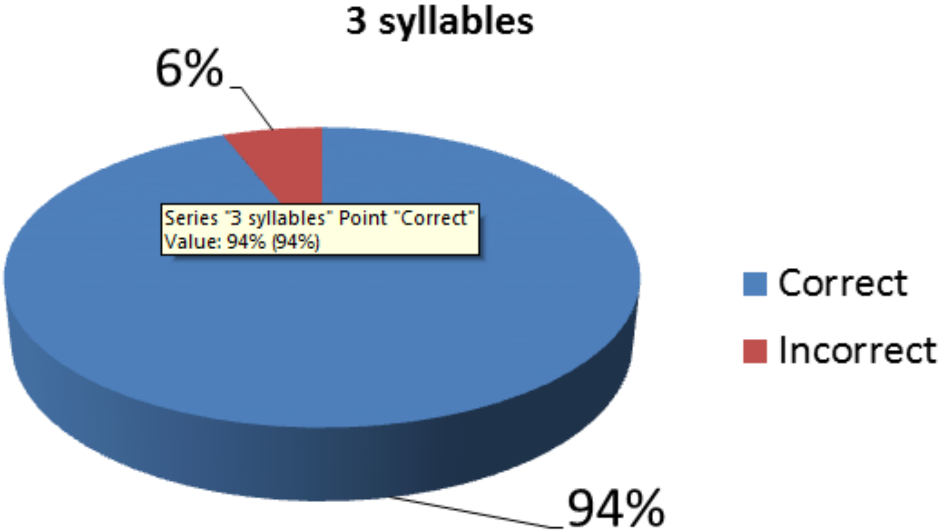
```
glm(formula = as.factor(correctness) ~ as.numeric(number_of_syllables) +
as.factor(last_sound), family = "binomial", data = mydata)
Coefficients:
                                  Estimate   Std. Error  z value Pr(>|z|)
(Intercept)                      1.717e+01  6.523e+03   0.003   0.9979
as.numeric(number_of_syllables)  1.396e+00  8.276e-01   1.687   0.0916 .
as.factor(last_sound)a          -1.841e+01  6.523e+03  -0.003   0.9977
as.factor(last_sound)b           1.060e-08  9.224e+03   0.000   1.0000
as.factor(last_sound)d          -1.787e+01  6.523e+03  -0.003   0.9978
as.factor(last_sound)g          -1.396e+00  9.224e+03   0.000   0.9999
as.factor(last_sound)k          -3.992e+01  9.224e+03  -0.004   0.9965
as.factor(last_sound)l          -1.857e+01  6.523e+03  -0.003   0.9977
as.factor(last_sound)n          -1.813e+01  6.523e+03  -0.003   0.9978
as.factor(last_sound)o          -1.996e+01  6.523e+03  -0.003   0.9976
as.factor(last_sound)r          -1.396e+00  9.224e+03   0.000   0.9999
as.factor(last_sound)s          -1.874e+01  6.523e+03  -0.003   0.9977
as.factor(last_sound)t          -1.759e+01  6.523e+03  -0.003   0.9978
```



1 syllable



2 syllables



3 syllables

	Number of syllables		
Correctness	1	2	3
0 (No)	4	5	1
1 (Yes)	10	15	15

Conclusion

- more syllables - better prediction
- grouping the features does not improve the outcome
- think about other factors

References

- Bauer, Laurie & Lieber, Rochelle. 2013. The Oxford guide to English morphology. Oxford: Oxford University Press, in press. PLAG, Ingo.
- Bernd, Heine & Claudi, Ulrike. 1991.G rammaticalization: A conceptual framework. Chicago: University of Chicago Press. 328 pages.
- Carstairs McCarthy, Andrew. 1992. Current morphology. Linguistic Theory Guides. London: Routledge. Pp. 289
- Coccaro, Noah & Jurafsky, Daniel. 1998. Towards Better Integration of Semantic Predictors in Statistical Language Modeling. In P roceedings of the
- International Conference on Spoken Language Processing (ICSLP98), Vol. 6, p. 2403–2406.
- Daelmans, Walter & Bosch, Antal. 2009. Memory-Based Learning. A draft chapter for the Blackwell Computational Linguistics and Natural Language
- Processing Handbook. ILK/Tilburg centre for Creative Computing, Tilburg University. Tilburg.
- Erben, Johannes. 1983. Einfuhrung in die deutsche Wortbildunglehre. 2 nd ed. Berlin:
- Schmidt Goldsmith, John & Riggle, Jason. 2011.The Handbook of Phonological Theory. 868 p.
- Jurafsky, Daniel. 1988. On the semantics of the Cantonese changed tone. Linguistics Society (BLS 14) , 304–318, Berkeley, CA.
- Leech, Geoffrey. 1983. Principles of Pragmatics. London: Longman .
- Schneider, Klaus P. 2003. Linguistische Arbeiten: Diminuitves in English. Tubingen: Max Niemeyer. 266 pages.
- Skousen, Royal. 1989. Analogical Modeling of Language. D ordrecht: Kluwer Academic Publishing.
- Wurstle, Regine. 1992. Uberangebot und Defizit in der Wortbildung. Eine knotrastive Studie zur Diminutivbildung im Deurschen, Franzosischen und Englischen. Frankfurt.M., etc.: Lang.