Productivity of the Broken Plural in Maltese

Shiloh Drake & Rebecca Sharp
6th International Conference of Maltese Linguistics
June 8-9, 2017

Introduction

Singular	Plural	English
ktieb	kotba	book/s
ġakketta	ġkieket	jacket/s
lupu	lpup	wolf/ves
kaxxa	kaxex / kaxxi	box/es

- Which features are important for inducing the use of a broken plural in Maltese?
  - Approximately 69% of Maltese nouns take a set of suffixes (the sound plural)
  - ~7% take a broken plural
  - Some take sound OR broken plurals

(Schembri, 2012)

Introduction

- Previously: Broken plurals cannot be predicted and are mostly random (Sutcliffe, 1936)
- More recently: Actually, the form of the broken plural *can* be predicted by the singular (Schembri, 2012; Mayer et al., 2013)
  - Schembri proposed 11 classifications based on the CV structures of the templates
  - Mayer et al. formed 5 classifications based on plural CV structure, moraic weight of the singular, and syllable structure of both forms
    - Class membership was also able to be predicted using this model

Previous Attempts

- Farrugia & Rosner (2008) used a neural network to predict the CV skeleton of a broken plural given the singular form
  - Problem: Neural nets tend to overfit models unless they have a HUGE amount of data
  - Problem: If overfitting occurs, they have trouble generalizing to new examples
  - Problem: Trained models are also uninterpretable, and thus unable to inform theories of language acquisition or language learnability

Our Solution?

- Using a generalized context model (GCM) (Nosofksy, 1990)
  - Similar model was used to predict Arabic broken plurals given a singular form (Nakisa et al., 2001; Dawdy-Hesterberg & Pierrehumbert, 2014)
  - This model attempts to classify new items based on their similarity to existing items
  - Especially appealing given human language generalizations based on analogy (Hay & Baayen, 2005; Bybee, 1995; Rumelhardt & McClelland, 1987)

Our Solution?

- Behavioral data then validates the model
    - A free response method similar to a Wug Test (Berko, 1958), where participants are asked to provide the plural for both real Maltese words and nonsense words
    - Used nonsense words that follow the CV skeleta predicted by the machine learning algorithm

Machine Learning

- Three models are compared
  - Baseline: GCM, which takes the number of items sharing a common pattern (lexical gangs) into account
  - *k*-Nearest-Neighbors (*k*NN) model, which is nonlinear and takes only individual lexical items into account
  - Logistic regression classifier, which learns weights over independent features and allows us to analyze the relative importance of each feature for predicting the form of the broken plural

Machine Learning

- Methods
  - Trained each model on 546 singular/broken plural pairs
  - Five-fold cross-validation: 3 folds for training, 1 for development, 1 for for testing
    - Any lexical gangs with fewer than 5 members were removed (109 members; original data set = 655 pairs)
    - All items were randomly balanced across the folds

Machine Learning

- Methods
  - Used both restricted and unrestricted GCMs
    - Restricted models classify test forms only to categories that match the CV skeleton of other singular forms the model was trained on, while unrestricted models classify test forms to any category
  - Logistic regression was used to measure the contributions of various aspects of similarity to the classification mechanism
    - Classification is given a degree of confidence as well: If the probability of choosing a certain classification is 90% as opposed to 60%, we know the classifier is more certain of its choice

Machine Learning

- Methods
  - Logistic regression was used to measure the contributions of various aspects of similarity to the classification mechanism

    1. Maximum similarity
       - Similarity between test item *i* and all items *j* in class *J*; predictive of class membership if test item is highly similar to at least one member in the class
    2. Average similarity
       - Similarity between test item *i* and all items *j* in class *J*; predictive of class membership if test item is highly similar to all members in the class

Machine Learning

- Methods
  - Logistic regression was used to measure the contributions of various aspects of similarity to the classification mechanism

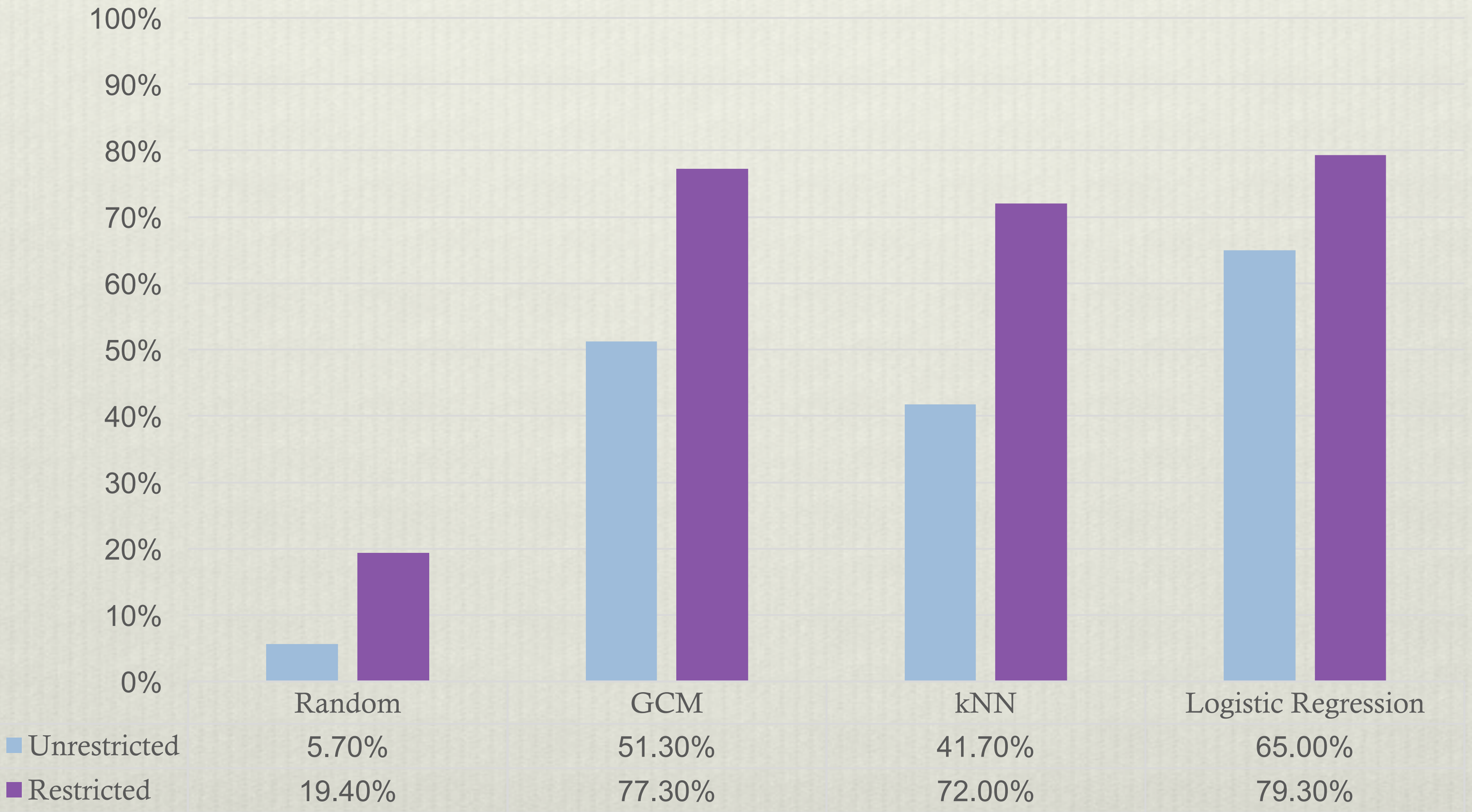
    3. Class size
       - Classes with more members are overall more probable, so are more likely classifications for the test item
    4. GCM similarity
       - Similarity between test item *i* and class *J*; combination of average similarity and class size

Machine Learning

- Results
  - In unrestricted models, logistic regression model performed better than both the GCM baseline and the *k*NN model ($p < 0.001$)
  - Restricted models were able to correctly classify broken plurals more often than the unrestricted models ($p < 0.001$)
    - When models are restricted, there are fewer choices for classification
      - Restricted models chose between 5 classes on average, while unrestricted chose between 18
    - Logistic regression model performed better than the *k*NN model ($p < 0.001$)
    - The logistic regression model performs somewhat better than the GCM, but the difference is not significant ($p = 0.06$)

Machine Learning

- Model accuracy



Machine Learning

- Discussion
  - Analysis of features from the logistic regression
    - Features that were independent of the size of the gang (maximum and average similarity) accounted for most of the performance
    - Size-dependent features were largely not orthogonal to to the size-independent features
    - When choosing only between classification options that share the same CV template as the singular in the test item, the relative size of the gang is more important
    - When all lexical gangs are options, the similarity between items matters the most

Machine Learning

- Discussion
  - Our GCM did not perform as well as Dawdy-Hesterberg & Pierrehumbert's for Arabic broken plurals
    - They had both sound and broken plurals and more than twice as many lexical gangs
    - Differences could be due to not tuning the hyperparameters of our model specifically to Maltese data, differences in model implementation, or differences in suitability of the GCM for Arabic and Maltese
  - While our LR model is not significantly more accurate than our GCM, it has advantages:
    - Fewer false positives
    - GCM seems to over-value gang size, while LR seems to value each factor equally

Machine Learning

- Discussion
  - Similarity of features between test items and classified items mattered most to the models, followed by the size of the lexical gang—what about in humans?
  - Experiment 2: A word elicitation task

Word Elicitation Task

- Methods
  - Participants (n = 14; age 18-49 (median 20.5); 12 RH; 6 female) provided the plural for the words that they saw
    - e.g., they were given “kotba” and wrote in the answer
  - 72 words total; 36 novel, 36 real
    - 18 real words took sound plurals, 18 took broken plurals
    - Novel words matched CV skeleta of the lexical gangs used by the models
    - Novel words did not necessarily have an easily parsed triconsonantal root
      - C.f. *lupu* => *lpup*; *ġakketta* => *ġkieket*

Word Elicitation Task

xewka	tifkiż	gemnilu	tqirra	tamdi	tirniċ
xwiek	tfafkiż	gmnili	tqirer	tamad	trineċ
xwieki	tfakaż	gmielen		timed	trieneċ
	tfiekeż	gmiemel		tmied	tirnuċ

- Participants responded using both sound and broken plurals
  - Sound plurals were more likely overall, but each nonsense word received responses that were broken plurals from at least two participants

Word Elicitation Task

- Discussion
  - Participants followed three main patterns given the CV structure of the singular:

1. CVCCV singulars typically had plurals with C_1C_2 cluster at the beginning and C_3 word-finally, leading to resyllabification as a single heavy syllable
   - tamdi → tmied, xewka → xwiek; żonta → żnut, toqxa → tqux
2. Alternately, CVCCV singulars could be resyllabified with glides or epenthesized vowels
   - żonta → żonot, xesna → xiesen, toqxa → toqox
   - xesna → xnejjes, toqxa → tqejjex
3. CVCVC singulars tended to have a CVCCV plural
   - naġat → naġtu

General Discussion

- While the modeling results and the human results are not terribly well correlated, ablation tests suggest that both the models and the humans use similar features to determine the form of a broken plural
  - When predicting whether a word will take a broken or sound plural, we can expect that the sound plural will be the default, but similarity to other words that take a broken plural can trigger the use of a broken plural
  - If a gang is larger, then at least the models will be more likely to select it with everything else being equal, and humans will likely do the same

General Discussion

- Krott et al.'s (2001) and Hay and Baayen's (2005) analyses of analogy in morphology may be helpful in interpreting the human results
  - Without concrete rules for selecting a broken plural, and with variation in how participants respond to a single nonsense word, they may be tapping into their lexicon for similar-sounding words and using the most frequent CV template
- This also fits with previous psycholinguistic literature on the importance of the CV skeleton in Maltese (e.g., Galea, 2011) and in Arabic (Dawdy-Hesterberg & Pierrehumbert, 2014; Boudelaa & Marslen-Wilson, 2004; McCarthy & Prince, 1990)

Next Steps

- Finish collecting human data
- More fine-grained analysis of the data with possible contributions from other factors in addition to gang size and similarity to other lexical items
  - e.g., biphone or triphone ngrams, CV-skeleta ngrams, word frequency, frequency of forms (overall and as plural morphemes)…

Thank you!

sndrake@email.arizona.edu / bsharp@email.arizona.edu