Talking Turkish: Using N-Grams for Syndromic Surveillance in a Turkish Emergency Department without the Need for English Translation

Sylvia Halasz PhD[1], Philip Brown[1], Cem Oktay MD[2], Arif Alper Cevik MD[3], Isa Kilicaslan MD[2], Colin Goodall PhD[1], Dennis G Cochrane MD[4,5], John R Allegra MD, PhD[4,5], Guy Jacobson PhD[1], Simon Tse PhD[1]

AT&T Labs – Research[1], Akdeniz Üniversitesi, Turkey[2], Eskisehir Osmangazi Üniversitesi, Turkey[3], Emergency Medical Associates of NJ Research Foundation[4], Morristown Memorial Hospital Residency in Emergency Medicine[5]

Introduction: Previously we used an "N-Gram" classifier for syndromic surveillance of emergency department (ED) chief complaints (CC) in English for bioterrorism. The classifier is trained on a set of ED visits for which both the ICD diagnosis code and CC are available by measuring the associations of text fragments within the CC (e.g. 3 characters for a "3-gram") with a syndromic group of ICD codes. Because the ICD system is language independent, the technique has the potential advantage of rapid automated deployment in multiple languages. Our objective was to apply the N-Gram method to a training set of Turkish ED data to create a Turkish CC classifier for the respiratory syndrome (RESP) and determine its performance in a test set.

Objectives: To determine how closely the performance of an ngram CC classifier for the Respiratory (RESP) syndrome matched the performance of the ICD9 classifier.

Methods: Design: Retrospective cohort. Setting: University hospital ED in Turkey. Participants: All ED visits in the database for 2002. Protocol: Two of the authors created by consensus a respiratory grouping of ICD10 codes chosen to be similar to a standard respiratory grouping of ICD9 codes created by the ESSENCE-CDC project. We then used an N-Gram method adapted from AT&T Labs' technologies applied to the first 10 months of data as a training set to create a Turkish CC RESP classifier. We next applied the classifier to the test set of visits for the last two months and determined the correlation for daily volumes measured by the CC classifier versus the RESP ICD10 grouping.

Results: The Turkish ED database contained 30,157 visits. The correlation coefficient was R =0.88.

Conclusions: The N-Gram method successfully created a CC RESP classifier in Turkish that performed similarly to the ICD10 RESP grouping. This was accomplished without translating the Turkish CC to English or understanding the Turkish language. This approach has promise in that it may offer a complementary method to using manual and natural-language techniques and has the advantages of systematic, consistent and rapid deployment as well as language independence