Noname manuscript No.

(will be inserted by the editor)

Ai Powered E-learning Esl System Of Vocabulary Acquisition For Beginners

Aleksey Sinyagin · Paola Di Maio

Received: date / Accepted: date

Abstract English language proficiency is increasingly necessary for acquiring a variety of competences and skills in most fields of knowledge. Due to shortage of teachers, especially in developing countries, the demand for English language skills training exceeds the supply. The availability of, and access to online English learning resources can open doors for self learning in countries where teacher population is scarce. We constructed a custom vocabulary made of 8,500 words from a corpus of children movies, cartoons and books, and conducted an experiment to illustrate various characteristics of the corpus and how such corpus can be used in game based ESL learning system. Based on the hypothesis that such characteristics, once identified and understood in the context of learning, could help to better develop educational programs such as evaluating the reading, comprehension and communication skills of ESL students, an experiment is conducted, where the students were presented the words from the custom developed English language vocabulary with the goal to guess the correct word translation. The words are clustered first by the frequency of appearance in the original corpus and later by the historical success of the words by the previous group of users. The experiment was run to detect if game based system can be used in ESL vocabulary development and acquisition and how the choice of words selected for beginner learners affects their engagement with the system. We conclude that vocabulary selection contributes to faster proficiency

Keywords

ESL, English, Learning, AI, automation, vocabulary acquisition, machine learning

Introduction

The English Language is increasingly important as the *lingua franca* for learning and education worldwide, especially in the technical and scientific domains [12] yet there is great shortage of qualified teachers: by 2030, at least 33 countries will still not have enough teachers to provide every child with a primary education. Unesco estimates that 25.8 million school teachers needs to be recruited to provide every child with a primary education, which include 3.2 million new posts and the replacement of 22.6 million teachers expected to leave the profession [1].

Motivation and Background

Although teachers are not likely to be entirely replaced by AI any time soon, educational and learning activities be supported by intelligent automated systems, and the landscape of platforms and services to support language learning is growing and becoming more specialised and diversified. Thanks to the advanced technical expertise of the team of volunteers of the Distance Teaching and Mobile Learning (DTML) a non-profit organization promoting English, advanced intelligent platform is offered at no cost to the public via the dtml.org website, Some advanced capabilities supported by machine learning have been developed and tested, the methodology and results presented in this paper

Hypothesis

The work presented in this paper is developed based on the hypothesis that am online game based platform can be used successfully in ESL vocabulary development and acquisition as the initial step of ESL self-learning process and that the order of words being introduced to the students play significant role in the learners engagement with the system and the student learning rate.

Research Design

As the first step of the experiment the custom vocabulary was constructed. The vocabulary was analyzed and each word was annotated with set of features derived from original corpus. The words were grouped into 10 buckets representing relative complexity of the word. Two different algorithms were used to group words into buckets. The online game was constructed to introduce to different words to students, and ask them to guess word translation (Figure 2). The game was published online and students across the globe were invited to play. In this particular experiment, which has taken

[1] http://www.teachersforefa.unesco.org/v2/index.php/fr/newss-2/item/490-global-teacher-shortage-threatens-education-2030

place from 26 of May 2018 to September 9th of 2018. During an experiment online ESL learning system was advertised on Google to attract ESL learners from different countries. There were no prior knowledge of user ability, location, native language, etc. The ads were generic and invited users to play games and learn english with the keywords such as “esl games”, “learn english”, “english games”. The ad ran globally across the world. The campaign generated 2.82M ad impressions on Google with 301,836 clicks. According to Google report 43,752 visitors were female and 80,634 were males, while the gender of 177,450 users was not identified. The majority of users visited the website on computers 212,849 while 84,608 users used their mobile phones and 4,379 users used tablets. The users came from total of 228 different countries with the top five counties being India, Vietnam, Turkey, Argentina and Colombia. The users landed on the homepage of the website and have an option to choose their activity out of 20 different games or bounce (close the browser window without taking any action). The average bounce rate for the users in that time period was 5.91%. Only users who engaged with Word Battle game were considered to be selected for the experiment 15,132 unique users .

For the testing methodology we used standard A/B test approach [14], the visitors who engaged with the game were divided into two separate groups: Group A and Group B by using cryptographically secure pseudorandom number generator [4]. The cookie were dropped on each user local computer to ensure that repeated visits will place the use into the same experimentation group. The cookie based approach has a limitation of being bound the the user browser, meaning that if later user used different browser to access the games, they can be assigned to a different experiment group. We assumed that percentage of such users is minimal and will not impact results of the experiment.

The students in group A were given the words in accordance to the frequency of the word appearance in the original corpus. If the user answers more frequent word correctly the less frequent word will be given. If user fails to answer the word correctly the less frequent word will be given. This approach operated under hypothesis that the more frequent the word is the more important it is to learn for the beginner users and it should be easier it is to be recognized and answered by the student. Student in group B were given words based on the historical success of learned by observing independent users playing the same game (Figure 2). The data was collected and analyzed based on two metrics: number of words answered by users correctly in each group, the length of the play and return rate of the users (how many times a particular user played the game within the observation period). The research methodology is outlined in Figure 1.

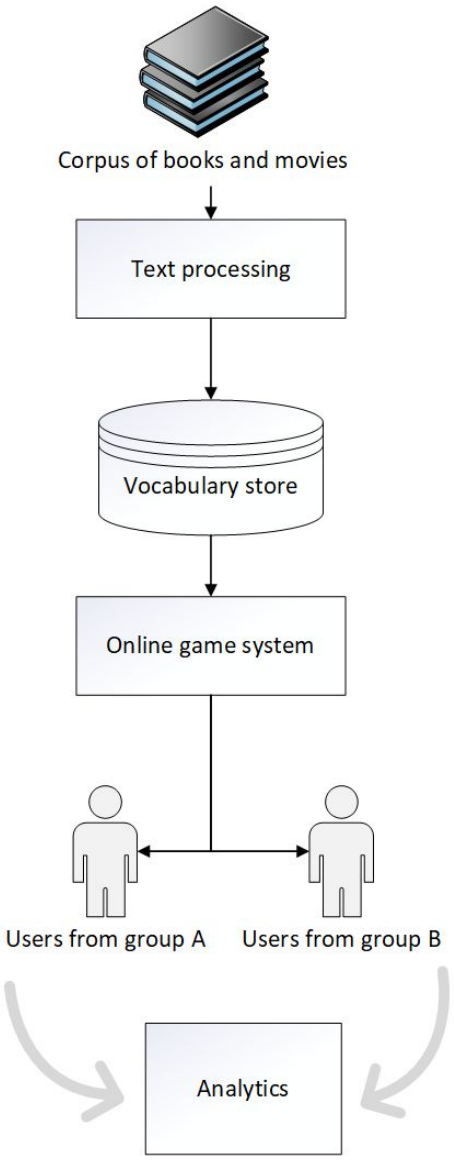

Fig.1 Research Design

Method

First, we analyzed a corpus of 48 documents including classics such as “the Wizard of OZ”, “Willy Wonka”, “Aladdin” and others. (ANNEX 1) The corpus was manually selected to contain G – General Audiences movies and books (All ages admitted. Nothing that would offend parents for viewing by children) in accordance to Motion Picture Association of America MPAA rating system [8].

Secondly, the corpus of documents was processed by applying tokenization [13] to extract all the unigrams (words). The simple approach to tokenization was used where

all non-alphanumeric characters were considered as a word boundary. For each unigram, we then applied stemming to obtain its corresponding stem. Stemming is a process to convert a word to its root and we used Porter- Stemmer [5]. Thus different words might be converted into the same stem. As different words might be converted into the same stem, the corpus of the documents contained 24,681 unique words with 17,609 unique stems. We then calculated the term frequency (TF) and the inverse document frequency (IDF) for all the tokens to understand frequency distribution of the words in the original corpus. IDF for a unigram shows the percentage of the documents that contain the unigram, thus indicates the popularity of the unigram in the corpus across all the documents. We computed a global TF and a document-wise TF. For a word t, its global TF is simply the number of times it occurs in all the documents in the corpus, its document-wise TF is the number of times it occurs in each document, its IDF is computed as IDF(t) = log(total number of documents in the corpus / Number of documents with term t in it). IDF equal to one means that the words was present in all documents in the analyzed corpus.We also calculated the frequency of each unigram in each document.

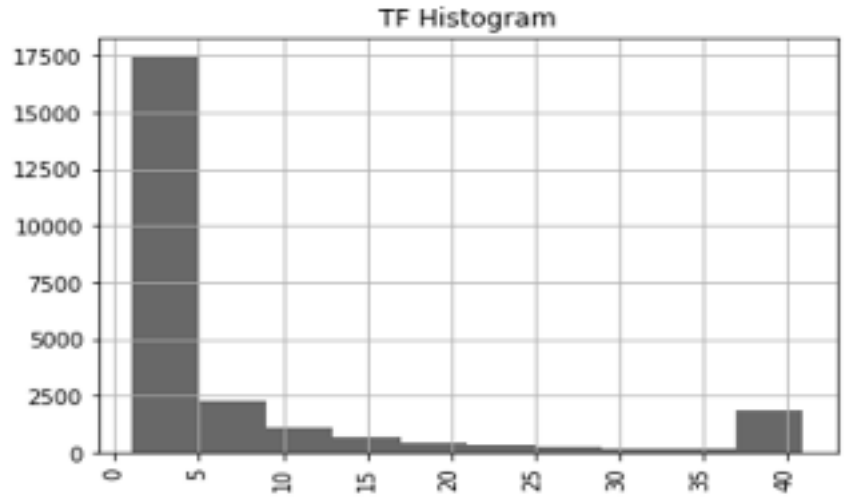

Figure 2. TF histogram for the word

We can see that most of the words have a relatively low term frequency (less than five) but there are still words having relatively high frequency (greater than 40). Notice we included all the words with frequency higher than 40 in the last bin and there are indeed some words with very high frequency such as more than 30000.Inverse document frequency (IDF) of each unigram shows the importance of the word across all corpus of

the documents.

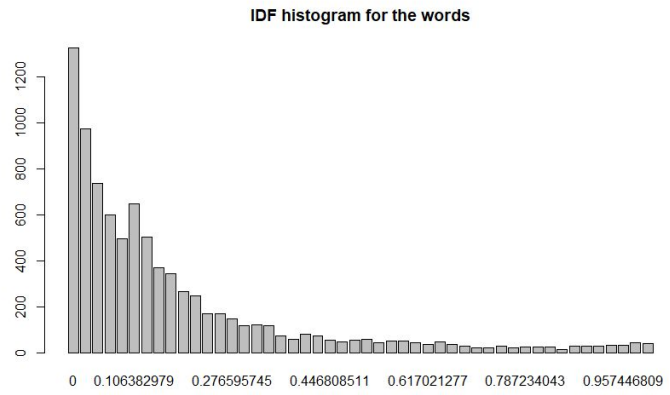

Figure 3. IDF histogram for the words.

IDF is calculated as logarithm of total number of documents divided by number of documents with word x. Similarly to TF, most of the words have a relatively low inverse term frequency (less than 0.1), namely they showed up in no more than 10% of the documents in the original corpus.

word	tf	idf
the	30272	1.00000000
to	13801	1.00000000
and	13272	1.00000000
you	12629	1.00000000
it	8273	1.00000000
in	6581	1.00000000
he	5714	1.00000000
on	5396	1.00000000
that	4436	1.00000000

Table 1. Example of top 10 words sorted by TF from extracted corpus.

The uni-grams were then passed through the profanity filter and 67 words not appropriate to children learning ESL were detected. The words excluded were “suicide”,

“fart” and others. [ANNEX 2]

Profanity filter is a list of 2320 words manually identified from English considered as strongly impolite, rude or offensive. Finally, the list of top words was extracted from the corpus. The words selected for this list met the criteria of appearing in at least three documents and having TF > 5. There were 8500 unique unigram words selected from the document corpus which we considered a base for teaching ESL. To understand if the words dictionary we build is an adequate dictionary to teach beginners level of english we conducted POS (Part-Of-Speech) tagging for every word in the corpus documents. We specifically tag people names as NAME and non-word token as NEW. For people names, we used the name corpus from python NLTK package [2]. Thus if a token is in the name corpus, we tag it as NAME. For non-word token, we used the English word dictionary from python Enchant package [3]. If a token is not in the English word dictionary, we tag it as NEW. For the remaining words, we applied python NLTK pos tagging [4].

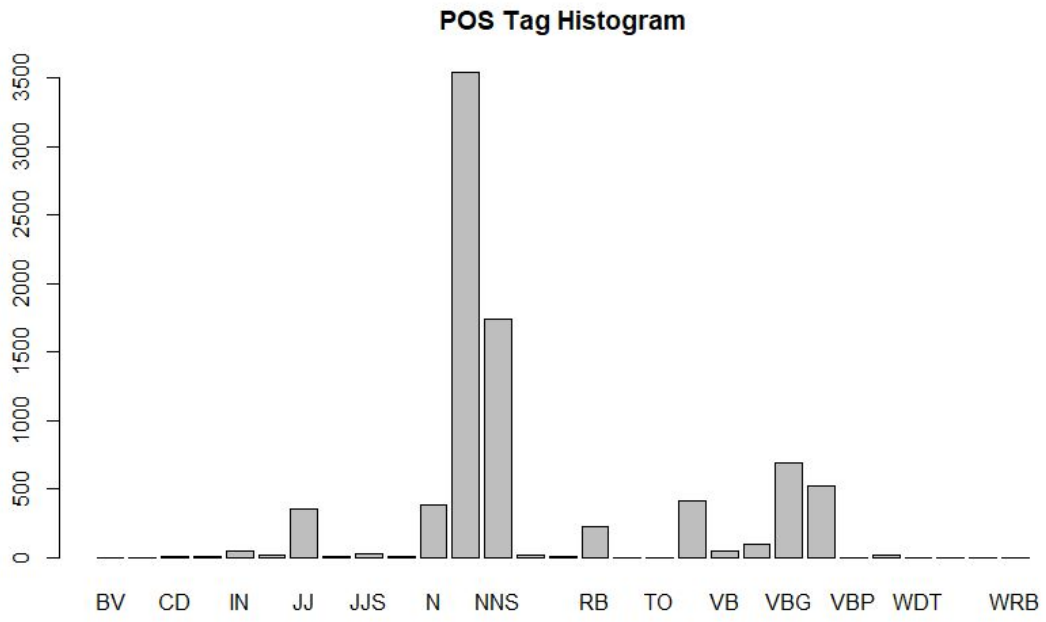

Figure 3. Plot the frequency of different pos tags over the corpus of documents

The plot of the frequency of different pos tags over the corpus of documents was constructed and we can see that Noun, plural, singular or mass, verbs and are adjectives are the most frequent POSs. Nouns and are much more frequent than the other pos tags. This distribution corresponds to the similar distribution of the learners’ corpus with N(Noun) and VB(Verb) are the first two leading POSs in both corpora [9]

One word can have multiple POS tags. Thus, we calculate for each word the percentage of each POS tag for the word. We also analyzed the distribution the number of tags associated with the words in the corpus. Most of the words in the corpus are associated with only one tag.

Finally, the words were grouped into 10 frequency groups based on words TF percentile distributions. Having constructed TF and TDF tables, frequency groups and POS tables, we were able to develop set of eLearning activities on DTML.org portal. By analyzing the success rate of students recognizing the words we can assess impact of the word selection algorithms on the students learning activity. To do so, we gave developed and experiment in the form of an English learning game called Words Battle. The game offers student English a set of words and ask them to translate that word into their native language. The translated word is double validated first against static, manually curated dictionary of translated words and if there is no match against online translation service to ensure translation is correct. Google translation API was used to re-validate the word in the case of miss-match with the local dictionary.

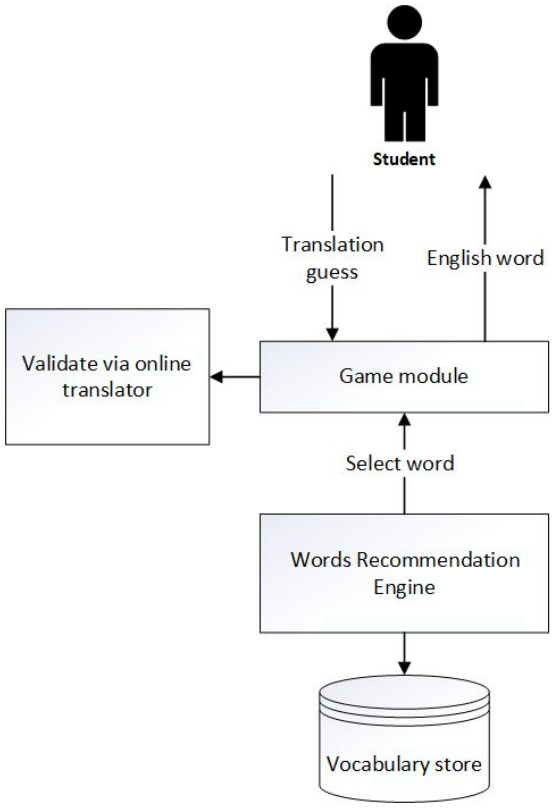

Fig.4 “Word Battle” Game structure

To collect baseline statistics, students playing on DTML.org portal wew of- fered randomly words from all 10 frequency groups . To ensure randomness the

cryptographically secure pseudorandom number generator RNG CryptoServiceProvider function was used. RNGCryptoServiceProvider is build upon Windows libraries for cryptographic operations, such as RSA and AES key generation [4].

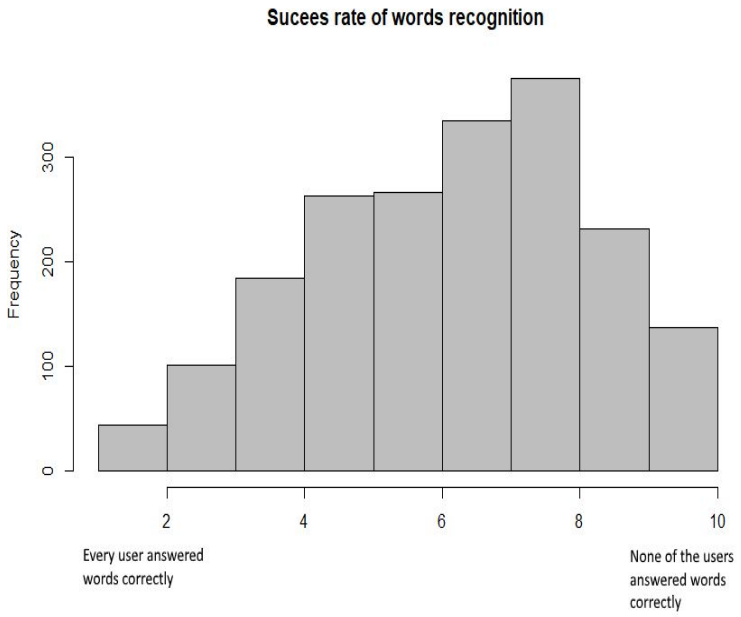

Fig 5. Words by success rate

Finally, the two algorithms were constructed one based on the term frequency and one based on the historical success rate as described in (Figure 2) and experiment was run for two groups of randomly selected users.

Results

As part of the experiment 15,132 unique users played the Words Battle games on the platform. With 7,780 users in the group A and 7,571 users in the group B. The group A students were presented the words based on the word frequency, as such as the most frequent words were show first assuming that frequency of the word indicates the importance of the word in the english dictionary and that word should be learned first. The group B students were offered words based on historical success of the word, no matter what the word frequency was. That approach focuses more on reinforcing student confidence by providing the words majority of students of a similar level answered correctly.

The collected data showed no correlation between success of the word and its frequency nor in children literature corpus nor in global internet corpus (Fig 6). Meaning the frequency of the words in the language corpus does not correlate with students

ability to recognize these words. This is interesting observation as it comes to design student curriculums for beginner learners.

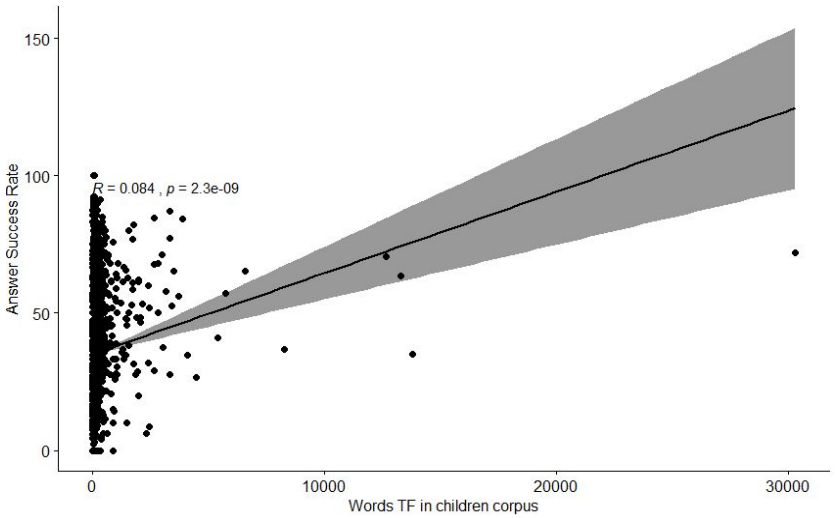

Fig.6 Pearson correlation between Word term frequence in children's corpus and students success rate of recognizing that word

Based on the observation above the which showed no correlation between success rate of recognizing the word and word term frequency, it was not surprising that users from the group B had better success in recognizing more words successfully

	Group A learners	Group B learners
Answered correctly	45,787	53,615
Answered incorrectly	63,582	60,112

Table 2. Word recognition comparison between two experimental groups of users

The Fisher statistical significance test showed p-value of both groups yielded the p-value < 2.2e-16 showing significance of the observed data. It also was interesting to observe that users from group B had longer engagement and more engaged learning sessions, whereby engagement is defined by the amount of time spent in session and number of words covered.

	Group A learners	Group B learners
Mean sessions length of active sessions (mi	316 mins	335 mins
Words coverage (Total number of unique words exposed to u each group)	2,400	2,590

Table 3. Engagement metrics for two experimental groups of users

Conclusion

The selection of the words for beginner ESL learners plays significant role in the way how students learn, engage and interact with online learning system. A system which have ability to dynamically select words for the learners in the way that students are both encouraged to learn from one side and challenged from another can improve the proficiency of learning. The approach above provides a simple metric for the evaluation ESL learning material selection and can be expanded to include more advanced methods of data analysis and expanded feature sets. As such, it would be beneficial to include POS and user native language characteristics as well as personal learning preferences.

References (to be edited, sorted)

[1] Davies, Mark. (2018-) The 14 Billion Word iWeb Corpus. Available online at https://corpus.byu.edu/iWeb/

[2] Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available online at https://corpus. byu.edu/coca/

[3] Davies, Mark. (2008-) Word frequency data, https://www.wordfrequency.info/iweb.asp

[4] RNGCryptoServiceProvider, Microsoft https://docs.microsoft.com/en-us/dotnet/api/system.security.cryptography.rngcryptoserviceprovide

[5] M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130-137.

[6] NetworkX, NetworkX.github.io [7] Distance Teaching and Mobile Learning, online gaming portal, https:/ /dtml.org

[8] "Classification and Rating Rules" (PDF). Classification and Rating Administration. January 1, 2010. pp. 6–8. Archived (PDF) from the original on December 4, 2014. Retrieved November 30, 2014

[9] Part-of-speech Sequences and Distribution in a Learner Corpus of En- glish , Proceedings of Research on Computational Linguistics Conference XIII (ROCLING XIII)

Taipei, Taiwan http://www.aclweb.org/anthology/O00-10 09

[10]RStudio Team (2015). RStudio: Integrated Development for R. RStu- dio, Inc., Boston, MA URL http://www.rstudio.com/.

[11] Alboukadel Kassambara, ggplot' Based Publication Ready Plots https://CRAN.R-project.org/package=ggpubr

We need reference to R Studio and packages we used Aleternative Target journal: https://www.igi-global.com/calls-for-papers/
international-journal-game-based-learning/41019

[12] Tatsioka, Z., Seidlhofer, B., Sifakis, N., & Ferguson, G. (Eds.). (2018). *Using English as a Lingua Franca in Education in Europe: English in Europe* (Vol. 4). Walter de Gruyter GmbH & Co KG.

[13] Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In *COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics* (Vol. 4).

[14] Online Controlled Experiments and A/B Testing,Encyclopedia of Machine Learning and Data Mining, pp.922-929, 2017 Ron Kohavi, Roger Longbotham

ANNEX 1 CORPUS (List the docs used)

Anastasia

The Little Mermaid

Willy Wonka

Sleeping Beauty

Frozen

Lion King

Wizard Of Oz

Tmnt

Garfield, The Movie

Bambi

Despicable Me

Moana

Aladdin

The Smurfs

Meet The Robinsons

Tangled

Hercules

Shrek

Despicable Me 2

Monsters Inc

Beauty And The Beast

A Bug's Life

Cars

Alvin And The Chipmunks

Cars-2

Toy Story

Rescuers-Down-Under

Shrek-The-Third

Finding Nemo

Goofy

Peter Pan

Incredibles

Mulan

Happy Feet

How To Train Your Dragon

Kungfu Panda

Wall-E

The Hunchback Of Notre Dame

Up

Brave

Toy Story 3

Coraline

Insideout

Fantastic Mr. fox

Antz

Cat In The Hat

How To Train Your Dragon

ANNEX 2 List of excluded terms