Input-Evaluation: A New Mechanism for Collecting Data Using Games with a Purpose

Adem Efe Gencer and Tunga Güngör
Department of Computer Engineering
Boğaziçi University
Bebek, 34342 Istanbul, Turkey
E-mail: efegencer@gmail.com, gungort@boun.edu.tr

Aslı Gürer and A. Sumru Özsoy
Program of Linguistics
Boğaziçi University
Bebek, 34342 Istanbul, Turkey
E-mail: {asli.gurer, ozsoys}@boun.edu.tr

***Abstract*—Collecting data through a game with a purpose (GWAP) has become a popular approach due to its numerous benefits. However, lack of diversity in game mechanisms puts some limitations on the areas of application of GWAPs. In this paper, we introduce a new two-phase mechanism for collecting data via human–based computation games. In the first phase, the players are provided with an object and asked to use that object within the domain of the game. In the second phase the players are provided with randomly selected data produced by the other players in the first phase and asked to evaluate them. A new game called "Dil Cambazı" that collects sentences containing passivized intransitive Turkish verbs using this mechanism is introduced and cases where input-evaluation mechanism is most useful are discussed.**

Keywords – Human Computation, Games With A Purpose, Input-Evaluation, Natural Language Processing

I. INTRODUCTION

The idea of game with a purpose (GWAP) was introduced by Luis von Ahn [1] in order to utilize human computation power via computer games. With this approach, due to their entertaining nature, people play games and as a side effect they solve some problems that are extremely challenging for computers. Each game employs a game mechanism, which determines the set of rules that are used in game design. Input and output of players, winning condition and the relationship between them are part of the game mechanisms. However, not every problem can be disguised in games using the existing game mechanisms. In this paper we introduce a new mechanism that enables us to create games for new problems.

Natural language processing (NLP) is a subcategory of computer science and linguistics. The ultimate purpose in this area is to have computers understand and respond to natural (human) language in the same way. Although the idea is fascinating, there are a number of hardships in the process of achieving this objective. In order to understand and analyze the semantic validity of a natural language text, computers should be taught the way humans acquire and use language, which is considered an AI-complete problem. Fortunately, contrary to computers, some of the tasks in these fields could be considered trivial for humans. Therefore, human-based computation is an extremely useful concept in such situations.

Determining whether a passivized intransitive Turkish verb could be used in the daily language and the semantic factors behind the choice is a problem that is very challenging for computers but trivial for Turkish-speaking humans. Although it is possible to passivize intransitive verbs in Turkish, when they are considered within the daily language, lack of semantic validity restricts the verbs that can in fact be used in such cases. A human can easily recognize the oddness of such verbs within sentences, but there is yet no feasible, straightforward algorithm for a computer to perform this task. For all these reasons, an approach that enables us to collect data from humans is developed.

In this paper, a game with a purpose which utilizes human computation power (www.dilcambazi.com) is introduced. The name of the game is "Dil Cambazı'-'Language Acrobat", which is a Turkish idiom denoting a person who uses the language in an extremely impressive manner. The game structure developed is a two-phase approach which utilizes a different strategy for each phase. Our initial game structure used a specialization of the output-agreement mechanism. At the end, data for creating a basic classification of intransitive verbs were collected. However, we observed that the initial game mechanism was not suitable for collecting data on semantic factors that lead to this classification. Moreover, the addictiveness rate of the game was not high enough to keep the rate of data collection high. We analyze the reasons behind the insufficiency of the output-agreement mechanism in the paper. Furthermore, we propose a new mechanism for collecting data using games with a purpose and discuss the areas where the new mechanism is suitable.

II. RELATED WORK

Although there are various games with a purpose in the literature, these games use one of the few game mechanisms. Three game mechanisms; output-agreement, input-agreement and inversion-problem were defined in [2]. Another mechanism, output-optimization, was defined in [3]. The idea of solving a difficult problem through a game was first

introduced by Luis von Ahn with ESP Game [4], which is one of the better-known games among all games with a purpose. ESP Game was a multiplayer game that pairs random players from all over the world to make them match their guesses on the same image within a limited amount of time. Time limit is used to enforce players to make their guesses as fast as possible, which helps extracting the first thought from the minds of players. ESP is the first game in which output-agreement game mechanism was used. As a result of the game, images on the internet have been labeled in a fast, accurate and cost-free way.

Output-agreement mechanism gained popularity and was used by a broad range of GWAPs. Matchin [5] is a two-player game that is based on showing two images to each player and asking them to choose the one that they think their partner would prefer. When they agree on the same image, they both earn points. Similarly, OntoTube [6], which is a game in OntoGame [7] project, followed the same approach to annotate videos using YouTube videos. SeaFish [8], which has the purpose of image labeling, and TubeLink [9], which is an improved version of OntoTube, are also part of OntoGame and utilize a mechanism that is a specialization of output-agreement games. This variance is due to the instant of output comparison. Contrary to the regular output-agreement games, the comparison is performed not just after each step, but at the end of each round.

Input-agreement [10] is another game mechanism, which is introduced with TagATune [11]. This two-player game is used for tagging music clips. Players listen to a music clip and describe it to their peer using tags. The aim is to find out whether the music clip they are listening to is the same or not. Unless both users agree on the correct answer, no point is earned.

In inversion-problem games, one of the players is designated as describer and the other player as guesser. The describer is given an object and provides hints to the guesser who tries to find out the object. If the object is identified, then the hints are considered to be descriptions related to that object. Verbosity [12] is a game that aims at collecting data about common-sense facts. One of the players provides certain properties about an object and the other player tries to find the object using the hints. Lebdemedenleblebi [13] is another inversion-problem game that is aimed at annotating a corpus. Peekaboom [14], which has the purpose of locating the characteristic region in an image, and Phetch [15], which has the aim of labeling images with explanatory descriptions to improve the accessibility of the web [16], used the same game mechanism.

Finally, output-optimization mechanism, which could be used for collecting subjective information [17], was used in games. In this mechanism, each player gets the same input and the outputs of players are given as hints to the other players. Restaurant game [18], which is a role-playing game aimed at collecting data on typical human behavior patterns, and Diplomacy game [19], which is a strategic board game that requires acting according to the behaviors of the others (i.e. creating and dissolving alliances), are the two examples of games that use this mechanism.

III. INITIAL GAME DESIGN

Passivization of intransitive verbs poses a challenge to NLP studies in Turkish. Although marking the intransitive verbs with the passive marker is generally productive in Turkish, the occurrence of some is restricted in sentences due to semantic invalidity. Our initial goal was to categorize intransitive verbs in Turkish as unaccusatives and unergatives. We used impersonal passive constructions as a syntactic diagnostic to categorize unaccusative and unergative verbs as compatibility in impersonal passive constructions indicates unergativitiy for the intransitive verbs [20], [21].

Unaccusative and unergative verbs differ from each other in the semantic and grammatical relations borne by the single argument of the verb. Unaccusatives are verbs whose argument is interpreted to be the theme argument in the object position. The single argument of the unergatives, on the other hand, is interpreted to be agent/experience subject of the verb. Unaccusatives and unergatives behave differently with respect to certain syntactic phenomena such as the impersonal passive construction. Their classification is hence crucial for NLP purposes.

In order to fulfill our goal, we designed the initial game with the aim of determining whether the players would accept the passivized version of an intransitive verb in the daily language. Players used simple radio buttons to answer the questions within a limited amount of time (Fig. 1). The game mechanism utilized a variant of output-agreement mechanism, in which the agreement of outputs was controlled using the answers to the seed questions. The seed questions were determined after a "base data collection" phase. In this phase, the game prototype was shared with a group of volunteers and answers were collected from them. Seed verbs were selected among the answers having a "Yes" or "No" answer rate above 75%. At each 240-second round of the game, 30 verbs were directed to the players. Players

Figure 1. Initial game interface

answered each question as “Yes (Evet)”, “No (Hayır)” or “Undecided (Kararsızım)” to see the next one. In each round, 5 out of 30 questions were selected among the pool of seed questions and the rest were randomly selected among the other intransitive Turkish verbs pool. To detect cheating, a strategy similar to the one used in [4] was followed by measuring the total time of answering questions. In cases of sharp decreases in total time, cheating is detected.

As a result of the game, intransitive Turkish verbs were located into a scale of unergativity versus unaccusativity. This scale was later used during the analysis of current game results. Contrary to most of the games in the literature, the initial game was not a browser-based game. This is a considerable advantage for players who do not have persistent Internet connection, but want to play the game. The result could be submitted anytime by the player using the “submit” button in the game. Players submitted their results in order to be a part of weekly high score table.

Although we collected enough data for creating a basic classification for intransitive verbs in terms of unergativity versus unaccusativity, semantic factors leading to this classification were still unknown. In order to understand the semantic factors, we needed to ask the users to provide us sentences using passive intransitive verbs which they considered to be a valid verb. Furthermore, the user feedback showed that lack of diversity in questions of the initial game caused a low addictiveness rate. These factors led to the creation of a new game mechanism, which enabled the players to use their full creativity, improve the in-game diversity and let us collect more comprehensive data: input-evaluation.

IV. THE NEW GAME MECHANISM

The initial game indicated that using output-agreement in our game design is not suitable for collecting data on semantic factors that determine the distribution of verbs within unergativity / unaccusativity scale. We needed a mechanism that would collect sentences containing the passivized intransitive verbs we provide. This mechanism had to collect high quality input and also keep the enjoyment rate of players high. To achieve these goals, we propose a new two-phase mechanism.

Our game mechanism is based on a single-player game. In the first phase (production-phase), a game object is given and the player is asked to use that object within the domain of the game. In this phase, no time limitation is imposed. The player is also encouraged to use items from a “recommended items list”, which would improve the quality of the produced output. In the second phase (evaluation-phase), the player is asked to evaluate randomly selected outputs of the other players from the first phase. The evaluation criteria are subjective. As a result, we not only collect subjective data, but also make people evaluate the subjective data we collect. We believe that our mechanism is particularly useful for NLP area and we introduce a game in which we employed the new mechanism in this area.

V. DIL CAMBAZI

In order to find out the semantic factors that play a role in the distribution of intransitive verbs within the unergativity / unaccusativity scale, we needed the users to construct sentences using the target forms. Moreover, we also needed to collect evaluations of the other players on these sentences to confirm the correctness of the input. Therefore, we employed the input-evaluation approach in the final game mechanism of Dil Cambazı. The interfaces of the first and second phases are hence different from each other (see Fig. 2).

To apply the new mechanism to our specific NLP problem in the first phase of the game the “game object” was selected among passivized intransitive Turkish verbs. In each round of the first phase, these verbs were given together with the root of the verbs, in order to ascertain that the players understand from which root the passivized verb is derived and construct sentences with these objects accordingly. Additionally, the “recommended items list” was constructed using words that would increase the semantic value of the sentences. The list was provided on the same game screen and titled under “Key”.

The first phase does not impose a time limit and consists of 18 rounds. In each round players either construct a sentence using the given verb or skip the question if they believe that no meaningful sentence can be constructed using the given passivized verb. To skip a question, a player simply hits the next button leaving the sentence field empty. While constructing a sentence, players are encouraged to use some words from the “Key” list. Right after the completion of the first phase (production-phase), the second phase (evaluation-phase) begins.

In the second phase, players are asked to evaluate the outputs that have been created in the first phase. These outputs could be classified under two types. The first type consists of sentences that have been constructed in the first phase. The second type consists of skipped rounds, indicating that no meaningful sentence was believed to be constructed using the provided passivized verb in that round. In the former case, the player (evaluator) is asked to evaluate whether the constructed sentence is meaningful and in the latter case the player is asked whether s/he agrees that no meaningful sentence can be constructed with the given intransitive verb. In both cases, the evaluations are made using radio buttons with label “Yes (Evet)”, “No (Hayır)” or “Undecided (Kararsızım)”. In a game session, a player can evaluate up to 30 answers before the time is up.

The most recent version of the game contains 273 intransitive Turkish verbs and their passivized forms. The verbs were selected among the ones that are still used commonly in the daily language. This lowers the probability of a player to encounter an unknown verb.

A. *Creating an Enjoyable Game*

One of the most important considerations in the game design is making people want to play it. Regardless of the

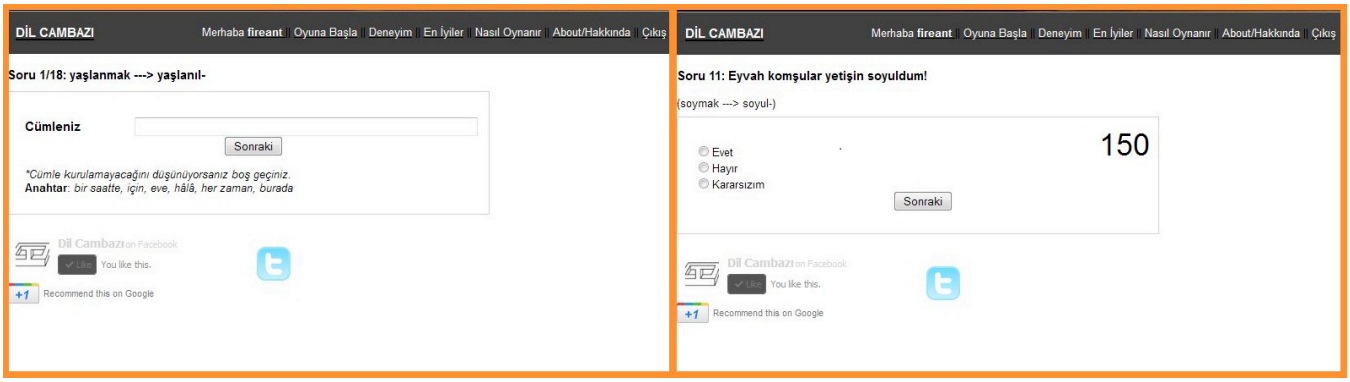

Figure 2. First phase and second phase interfaces (respectively)

game mechanism, creating a GWAP that cannot attract any player is worthless. If people enjoy the game then they would be willing to dedicate their time to it. And the more time people spend playing the game, the more user-generated data would be collected.

As mentioned in [22], [23], challenge is a crucial aspect of every computer game. In order to make the game more challenging and increase the addictiveness rate of the game, methods like time limitation, online score keeping, high-score list, randomness and various skill levels for players are adopted. However, while using these methods, it is crucial to keep the balance of difficulty and to design a game which is neither extremely challenging nor so easy.

Time limitation is used during the evaluation phase of Dil Cambazı. This limitation not only made the game more challenging, but also enabled players to evaluate each answer based on their initial reaction. This is especially an important criterion while collecting subjective information from players. In order to fully utilize the time limitation, it should be calibrated so that it is neither too short to make the game stressful nor so long to lose its meaning. Moreover, during an evaluation step a player makes a decision between a positive and a negative feedback. Lack of calibration in time limitation might lead to suboptimal decisions [24]. In the second phase of Dil Cambazı, a 240 seconds time limitation is imposed for 30 rounds. This limitation was calibrated according to the answer rates in this section and player feedbacks.

Online score keeping makes a game more addictive. People try to collect points and improve their score which is an indicator of the effort they put into the game. There are two types of scoring mechanisms. The first type is *instant scoring*, the second *brewed scoring*. In instant scoring, players earn points just after they complete a round. During the first phase, upon completion of a round a player earns 2 points. For the second phase, completion of each round brings 1 point. The brewed scoring mechanism works differently. Via this mechanism, players earn or lose points based on the other players' evaluations. In Dil Cambazı, a player earns 1 point for each positive evaluation of his/her sentence. On the other hand, he loses 1 point for each negative evaluation. Brewed scoring mechanism motivates players to enter correct input and helps prevent cheating by penalizing the abusers.

Similar to score keeping, high-score list also encourages players to spend more time playing the game. High-score list contains the most successful players in that game's history. In Dil Cambazı, 10 players having the highest scores are part of this list. Their ranking, username, total points and levels are listed in the high-score list.

Randomness improves the variety within the game. Dil Cambazı is very rich in randomness. During the first phase of the game, a small number of verbs are randomly selected among all the verbs in the database. The real randomness is observed in the second phase of the game, in which user-constructed sentences are selected among a growing number of sentences. One significant aspect of the game is that as long as players contribute to the game with new sentences, it will never lose its variety.

Representation of various skill levels is also a very efficient method of increasing the fun factor in a game. Players have various skill levels according to their total scores. The level names are not selected among commonly used names like "beginner", "expert", etc. Instead, fun names are used to further improve the fun factor. For instance, one of the level names, "Tosun Pasa", was selected from the character name of one of the highest rated Turkish comedy movies in IMDB database [25]. The names of the levels are not declared beforehand, which increases the curiosity level of the players and encourages them to play more to see the next level name.

In addition to in-game fun mechanisms, we provide some fictional articles for entertainment purposes on the home page. Furthermore, we incorporated functionalities from popular social networks (Facebook, Google+ and Twitter) to expose our game to a broader audience.

VI. RESULTS

A. *Initial Game Results*

In order to provide platform independency, the initial offline game was implemented in Java and was hosted in a Wordpress blog. The initial game-play period continued for 2 months and as a result categorization of intransitive verbs

in Turkish as unaccusatives and unergatives was completed. In total 273 intransitive Turkish verbs were categorized. The results of the initial game re-confirmed the 5-way distinction on unaccusative / unergative scale proposed by [20].

B. Final Game Results

The final online game was implemented in PHP using the Zend Framework. The game was hosted on a publicly accessible web site. It was released on July 28, 2011 and game-play period continued for 5 months. Semantic factors that play a role in the distribution of intransitive verbs in unaccusativity / unergativity scale were collected using the constructed sentences.

During the game-play period, in total 1031 sentences were constructed in the first phase. Moreover, 1262 evaluations were performed in the second phase of the game. The results show that, 73.6% of all evaluations were answered positively, 21.8% of the evaluations were answered negatively and remaining 4.6% of the evaluations are answered as undecided. This indicates that people generally make a positive or negative evaluation and majority of the evaluations seem to be positive which is in parallel with our expectations.

The second phase of the game aimed at determining whether semantic factors have an effect on the scalar distribution established in the first phase. Animacy of the internal argument is stated to be crucial in determining the syntactic behavior of verbs in Turkish [26]. Our findings also reconfirmed that when the implicit argument is interpreted to be animate, verbs which inherently take both animate and inanimate arguments were acceptable in the impersonal passive form, thus behaving as unergative.

Similarly, when their implicit argument was interpreted to be animate, verbs of manner of motion and change of location which typically allow both animate and inanimate arguments exhibited unaccusative properties in phrases with a telic interpretation indicating directed motion or an endpoint.

In our study the intransitive forms are used in impersonal passive form which is compatible only with animate arguments. Further focus on structures with intransitive verbs taking inanimate arguments will contribute to a more comprehensive analysis of intransitive verbs in Turkish.

VII. CONCLUSION AND FUTURE WORK

Although several games with a purpose have been created to solve problems in various fields, the mechanisms used to design those games have not changed much. This lack of diversity limits the possible areas in which human-based computation games could be utilized. In this paper we introduced a new game mechanism, input-evaluation, that is particularly useful for collecting subjective data.

We developed a new game which employs the input-evaluation mechanism to collect data on semantic factors that lead to the classification of intransitive Turkish verbs within unergativity / unaccusativity scale. We believe that the mechanism introduced in this paper will help broaden the areas of application for collecting data using games with a purpose.

As for future work, we are planning to develop a special game mode that is suitable for mobile devices. Since text entry using a mobile device would be a cumbersome process, a new game mode for such devices would be helpful for improving the playability. In this mode, the requirement of text entry is planned to be removed. Moreover, we are aiming to increase the amount of skill levels in Dil Cambazı. We hope that this update would improve in-game diversity and attract more players.

ACKNOWLEDGMENT

This work was partially supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) BİDEB under the programme 2219.

REFERENCES

[1] L. von Ahn, “Games with a purpose,” *IEEE Computer*, 39(6):92–94, June 2006.

[2] L. von Ahn and L. Dabbish, “Designing games with a purpose,” *Communications of the ACM*, 51(8):58–67, August 2008.

[3] Man-Ching Yuen, Ling-Jyh Chen, and Irwin King, “A survey of human computation systems,” *IEEE Symposium on Social Computing Applications (SCA'09, in conjunction with IEEE SocialCom'09)*, Vancouver, Canada, 2009.

[4] L. von Ahn and L. Dabbish, “Labeling images with a computer game,” in *Proceedings of the SIGCHI conference on Human factors in computing systems* (Vienna, Austria: ACM, 2004), 319-326.

[5] Severin Hacker and L. von Ahn, “Matchin: Eliciting user preferences with an online game,” in *ACM Conference on Human Factors in Computing Systems, CHI 2009*. Pages 1207-1216.

[6] Katharina Siorpaes and Martin Hepp, “Games with a purpose for the semantic web,” *IEEE Intelligent Systems*, Vol. 23, No. 3, pp. 50-60, May/June 2008.

[7] Katharina Siorpaes and Martin Hepp, “OntoGame: weaving the semantic web by online gaming,” in *Proceedings of the European Semantic Web Conference (ESWC)*, Springer LNCS, Tenerife, Spain, June 2008.

[8] Stefan Thaler, Katharina Siorpaes, David Mear, Elena Simperl and Carl Goodman, “SeaFish: a game for collaborative and visual image annotation and interlinking,” in *Demo Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*, Springer LNSC, Heraklion, Greece, May 29th - June 2nd, 2011.

[9] TubeLink. http://ontogame.sti2.at/games/.

[10] E. Law and L. von Ahn. “Input-agreement: A new mechanism for data collection using human computation games,” in *ACM CHI*, 2009.

[11] Edith L. M. Law, L. von Ahn, Roger B. Dannenberg, Mike Crawford, "Tagatune: a game for music and sound annotation," in *Proceedings of the 8th International Conference on Music Information Retreival*, Vienna, Austria, 2007.

[12] L. von Ahn, M. Kedia, and M. Blum, “Verbosity: A game for collecting common-sense facts,” in *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[13] O. Güngör, T. Güngör, “Morphological annotation of a corpus with a collaborative multiplayer game,” in *11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2010)*, Ed. A. Gelbukh, March 2010 Iaşi, Romania – LNCS (Lecture Notes in Computer Science), Vol.6008, 2010, p.74-85, Springer-Verlag, Berlin Heidelberg.

[14] L. von Ahn, M. Kedia, M. Blum, “Peekaboom: a game for locating objects in images,” in *CHI ’06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, New York, NY, USA, ACM pp. 55-64, 2006.

[15] L. von Ahn, S. Ginosar, M. Kedia, and M. Blum. “Improving image search with Phetch,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[16] L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum, “Improving accessibility of the web with a computer game,” in *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[17] Kam Tong Chan, Irwin King, Man-Ching Yuen, “Mathematical modeling of social games,” *CSE* (4), pp. 1205-1210, 2009.

[18] J. Orkin and D. Roy, “The restaurant game: Learning social behavior and language from thousands of players online,” *Journal of Game Development*, 3(1):39–60, December 2007.

[19] A. Krzywinski, W. Chen, and A. Helgesen, “Agent architecture in social games the implementation of subsumption architecture in diplomacy,” in *AIIDE*. The AAAI Press, 2008.

[20] M. Nakipoğlu-Demiralp, “The referential properties of the implicit arguments of impersonal passives in Turkish,” *Linguistik Aktuell (Linguistics Today)* 44. Amsterdam: John Benjamins, 2001.

[21] David M. Perlmutter, "Impersonal passives and the Unaccusative Hypothesis," in *Proc. of the 4th Annual Meeting of the Berkeley Linguistics Society*. UC Berkeley. pp. 157–189, 1978.

[22] T.M. Malone, “Heuristics for designing enjoyable user interfaces: Lessons from computer games,” in *Proceedings of the Conference on Human Factors in Computing Systems* (Gaithersburg, MD, Mar. 15–17). ACM Press, New York, pp. 63–68, 1982.

[23] T.M. Malone, “What makes things fun to learn? Heuristics for designing instructional computer games,” in *Proceedings of the Third ACM SIGSMALL Symposium and the First SIGPC Symposium on Small Systems* (Palo Alto, CA, Sept. 18–19). ACM Press, New York, pp. 162–169, 1980.

[24] Dan Ariely, and Michael I Norton, “From thinking too little to thinking too much: a continuum of decision making,” *Wiley Interdisciplinary Reviews Cognitive Science* 2, no.1, pp. 39-46, 2011.

[25] Tosun Pasa. http://www.imdb.com/title/tt0253828/.

[26] A. S. Özsoy, "Argument structure, animacy, syntax and semantics of passivization in Turkish: A corpus-based approach," in *Corpus Analysis and Variation in Linguistics*, Kawaguchi, 2009.