# The Importance of Visual Context Clues in Multimedia Translation

Christopher G. Harris[1] and Tao Xu[2]

[1] Informatics Program, The University of Iowa, Iowa City, IA 52242 USA
christopher-harris@uiowa.edu
[2] School of Foreign Languages, Tongji University, Shanghai 200092 China
michaeltjxt@hotmail.com

**Abstract.** As video-sharing websites such as YouTube proliferate, the ability to rapidly translate video clips into multiple languages has become an essential component for enhancing their global reach and impact. Moreover, the ability to provide closed captioning in a variety of languages is paramount to reach a wider variety of viewers. We investigate the importance of visual context clues by comparing transcripts of multimedia clips (which allow transcriptionists to make use of visual context clues in their translations) with their corresponding written transcripts (which do not). Additionally, we contrast translations produced using crowdsourcing workers with those made by professional translators on cost and quality. Finally, we evaluate several genres of multimedia to examine the effects of visual context clues on each and demonstrate the results through heat maps.

## 1 Introduction

Multimedia content-sharing websites invite global users to discover and share original video content. These websites have become immensely popular as demonstrated by their rapid growth; one such website, YouTube, has quickly grown to be the world's most popular video site since its introduction in early 2005. Each day over a billion video plays are initiated by millions of users on YouTube and similar websites [1]. These video content-sharing websites provide a forum for people to engage with multimedia content globally and also act as an important distribution platform for content creators.

However, their global reach is often limited by restrictions on their ability to translate into other languages by overdubbing or closed captioning. Original content creators are often individuals or small groups without access to substantial capital, limiting the opportunities to have their multimedia creations translated professionally. Also, with so many content creators vying for the limited time and attention of potential viewers, competition is keen; therefore, translations need to be performed quickly so as to capture and enhance international viewer interest before it wanes.

Currently there are several viable methods to translate multimedia content, three of which we explore in this paper. One method is to hire a professional translator to translate into several languages; however, this is costly and thus impractical for most contributors. Another method is to use machine (MT) translational tools, such as

P. Forner et al. (Eds.): CLEF 2011, LNCS 6941, pp. 107–118, 2011.

Google Translate, but the quality of these tools is not yet high enough to provide translations for complex concepts in other languages correctly. A third method is to obtain translation through the use of crowdsourcing, which in theory, permits translations to be performed quickly, correctly, and relatively inexpensively. In this paper, we will use the professional translation as our gold standard and explore MT and crowdsourcing approaches.

Evaluating these approaches raises some important issues on the use of visual context cues in multimedia translation. First, is it sufficient to work from a written transcript, as MT tools are required to do, or are the visual context cues found in video truly beneficial for translation quality? Second, since crowdsourcing makes use of humans who can make use of these visual cues, how does crowdsourcing compare to the MT tools available today? Third, how dependent are the previous two questions on the genre of multimedia we choose to translate?

The paper is organized as follows: In the next section, we discuss related work in this area. Section 3 contains a discussion of the Meteor MT evaluation system. In Section 4, we discuss our methodology and experimental approach. In Section 5, we present and discuss our findings and their resulting implications. We conclude and discuss directions for future work in Section 6.

## 2 Background and Motivation

The use of visual context as an aid to understanding is well covered in the literature. A number of visual contextual studies, (e.g.[2, 3]) have been conducted in the field of computational psycholinguistics, an area of linguistics concerned with the development of computational models that examine how language processing occurs in the brain. In addition, studies involving the utility of visual contextual cues have been explored for their effects on learning and attention [4-7], listening comprehension [8-10], speech and language comprehension [11-13], and second language retention [14-16]; in each of these studies, clear benefits of visual contextual clues have been identified.

Multimedia translations are normally straightforward extensions of the above, and therefore limited research has been focused on the use of visual context; however, some studies have found that translations involving creative imagery can possibly mislead [17] or unintentionally misinform [18]. In addition, if a user unfamiliar with a language observes imagery that is intended to be confused with the regular context of that word (i.e., a “play on words”), it is understandable how visual contextual clues can hinder, not aid, language comprehension as indicated in [19]. Therefore, we examine the use of imagery to examine this aspect.

Most machine translation (MT) tools work by using linguistic rules, using corpus statistics, using examples, or a combination of these techniques to determine the correct inter-lingual substitution between words or phrases. Many of these MT tools are freely available on the internet, such as Google Translate[1], Bing Translator[2] and Babelfish[3]. Although acceptable for simple translation tasks, the quality of MT tools

[1] translate.google.com
[2] microsofttranslator.com
[3] www.babelfish.com

is still too poor to use in a professional setting [20]. In fact, since the early 1960s, several doubts have been expressed about the ability to ever achieve fully-automated MT of high-quality [21] and that perfect translations would never be achievable [22].

We also investigate translations using crowdsourcing. Since its introduction in 2006, crowdsourcing has become a viable platform for the "crowd" - a large pool of semi-anonymous users - to perform a set of structured tasks [23]. Crowdsourcing marketplaces, such as Amazon's Mechanical Turk[4] are designed as a labor clearinghouse for "micro-tasks" – small tasks that can be done by anyone who meets preset qualification criteria – in return for payments based on number of tasks completed. Crowdsourcing focuses on micro-tasks requiring human intelligence, and therefore is applicable to the field of translation: work done by crowdsourced workers can be accomplished quickly, inexpensively, and has been demonstrated to be good quality [24], particularly if these micro-tasks are clearly defined and multiple participants perform the same task as a quality check [25].

We examine the use of crowdsourcing in this study since it provides an inexpensive yet reliable substitute for professional translation services [26]. Indeed, a recent study found that crowdsourcing speech transcriptions was nearly as reliable as professional translations but at 1/30$^{th}$ the cost [27]. More importantly, crowdsourcing allows us to examine the role of visual context clues in multimedia translation, which we cannot do with online MT tools.

## 3 The Meteor MT Evaluation System

We now turn our attention to the evaluation tools for translations. The evaluation tool we use in this study, Meteor [28], was introduced in 2004 to tackle some of the issues other MT evaluation systems did not adequately address. Meteor was designed to improve correlation with human judgments of MT quality at the segment (sentence) level; it has been shown to correlated better with human judgments than other MT systems [28]. Meteor evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a given reference translation. If more than one reference translation is available, the translation is scored against each reference independently, and the best scoring pair is used. Alignments are built incrementally in a series of stages using the following Meteor matchers:

- *Exact*: Words are matched if and only if their surface forms are identical
- *Stem*: Words are stemmed using a stemmer, such as the Porter Snowball Stemmer [29] and matched if and only if the stems are identical.
- *Synonym*: Words are matched if they are both members of a synset (synonym set) according to the WordNet database [30]. This ability to use synonym sets is powerful, since the choice of words used by two human translators may be very similar in meaning but not exact. This use of synsets allows for some flexibility in translation that is reflected in the real world. We made extensive use of this feature in our studies.

[4] mturk.com

At each stage, one of the above subroutines locates all possible word matches between the two translations using words not aligned in previous stages. An alignment is then identified as the largest subset of these matches in which every word in each sentence aligns to zero or one words in the other sentence. If multiple such alignments exist, the alignment is chosen that best preserves word order by having the fewest crossing alignment links. At the end of each stage, matched words are marked so that they are not considered in future stages. The resultant Meteor alignment used for scoring is defined as the union of all stage alignments.

The Meteor score for a given pairing is computed based on the number of mapped unigrams found between the two strings, *m*, the total number of unigrams in the translation, *t*, and the total number of unigrams in the reference, *r*. Unigram precision is calculated as $P = m/t$ and unigram recall as $R = m/r$. An F-measure, which is the harmonic mean of precision and recall, is then computed [31]:

$$F_{Mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

The value of $\alpha$ determines the tradeoff between precision and recall. The precision, recall and $F_{Mean}$ are all based on single-word matches, but the extent to which the word order matches also needs to be considered. Meteor computes a penalty for a given alignment in the following manner. First, the sequence of matched unigrams between the two strings is divided into the fewest possible number of chunks, maximizing the adjacency of matched unigrams in each string and in identical word order. The number of these chunks, *ch*, and the number of correct matches, *m*, is then used to calculate a fragmentation fraction = *ch/m*. To illustrate, a candidate translation that is an exact match with the reference document will result in a single chunk. The penalty is then computed as:

$$Penalty = \gamma \cdot (ch/m)^{\beta}$$

Here, the value of $\gamma$ determines the maximum penalty $(0 \leq \gamma \leq 1)$. The value of $\beta$ determines the functional relation between fragmentation calculated and the penalty. In practice, we empirically determine the optimal values for $\alpha$, $\beta$, and $\gamma$ for each language independently.

Although a number of MT evaluation systems exist and have their merits, we chose Meteor for a number of reasons. First, a number of studies have examined Meteor's correlation with human judgments across a number of scenarios. Additionally, we believe Meteor's use of synonyms provides some flexibility in capturing the essence of a translation better than some of the other metrics. Finally, the Meteor source code is well-maintained and readily available, permitting us to adapt the code, in particular the WordNet-based synonym module, to our specific needs.

## 4 Experimental Design

We used nine multimedia videos; each was considered challenging to translate due to the amount of figurative language they included. Our goal is to observe the effects of several different features on translation quality. We designed our experiments with

four separate features: multimedia genre, translation type, language, and whether or not visual context clues were used in translation.

Meteor values used for α, β, and γ for scoring each of the languages is provided in Table 1. They are optimized based on existing research [32] and from our own preliminary studies.

**Table 1.** Parameters used with Meteor

| | English | Spanish | Russian |
|---|---|---|---|
| α | 0.95 | 0.90 | 0.85 |
| β | 0.50 | 0.50 | 0.60 |
| γ | 0.45 | 0.55 | 0.70 |

We evaluate two hypotheses: first, the use of visual context clues provides an improvement in the Meteor evaluation score compared with using the written transcripts alone. Second, translations provided by crowdsourcing workers can obtain Meteor assessment scores as high as those from professional translators. If both hypotheses are true, this bolsters our overall claim that visual context clues matter in multimedia translation, and that accurate translations can be made at low cost.

**Table 2.** Mean Document Size, in words, by Language and Multimedia Genre

| | Chinese | English | Russian | Spanish |
|---|---|---|---|---|
| TS | 1,219 | 1,310 | 1,164 | 1,369 |
| AN | 684 | 705 | 624 | 743 |
| MV | 183 | 198 | 149 | 215 |

We tested our hypotheses as follows. We took nine short videos, each 5-10 minutes in duration, in Mandarin Chinese from three different genres – three talk shows (TS), three animated comedy skits (AN), and three music videos (MV). We chose multimedia videos with highly-figurative language content to ensure challenging translations. From these, we created written transcripts of each video clip in Simplified Chinese. Next, we hired three professional translators to provide transcripts in three different languages – English (EN), Spanish (ES) and Russian (RU). Figure 1 gives an illustration of the steps taken to break this translation study into several groups for each genre.

We conducted two separate studies – one using the raw multimedia to obtain visual cues while others used the transcripts only – in order to compare the difference in Meteor score. These translations were conducted using crowdsourcing (CS), online machine translation tools (MT) and professional translators (PT) – our gold standard. For the crowdsourcing translations, we hired non-professional translators through several crowdsourcing platforms to provide translations from Chinese into our three target languages. We took steps to ensure the same translator was not used to translate both from the multimedia and from the written transcripts alone for the same language pair, as this could introduce bias.

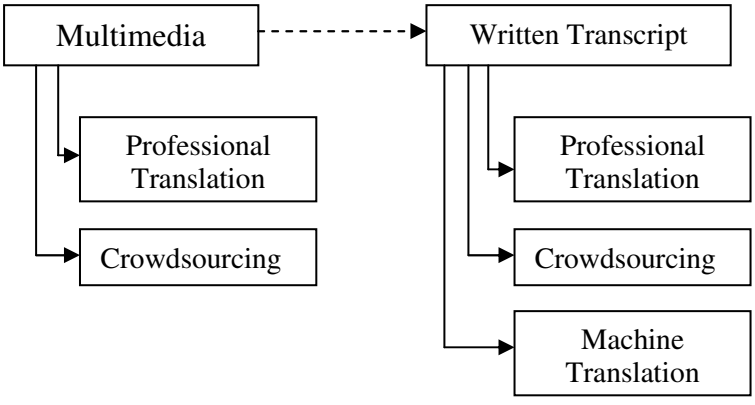


**Fig. 1.** Overview of the different groups evaluated in this paper

We also use the three aforementioned online MT tools from Google, Babelfish and Bing, as well as two others: Worldlingo[5] and China-based Lingoes[6] to provide translations from the written transcripts into our three target languages. We used the maximum Meteor MT evaluation score obtained from all five online translation tools for a single translation for a single genre. For crowdsourced transcripts, we had a minimum of two translations for each (with an average of 3.8 translations per transcript) and used the maximum score of these in our calculations. We then scored each using Meteor against our (gold standard) professional translation (PT) using the parameters given in Table 1. The version of Meteor we used for our evaluation includes support for English and Spanish. For Russian, we modified the Meteor program a Russian WordNet[7] into our Meteor system for synonym evaluation.

## 5 Results

Our first hypothesis examined whether the use of visual contextual clues improve the Meteor scores compared with the use of written transcripts alone. When evaluating our results (columns 1-3 and 4-6 in Figure 2), we are unable to conclude that the difference is significant when comparing the group means at $p = 0.05$; however, when we use a paired t-test, we are able to verify a significant difference at the $p = 0.05$ level of confidence.

Our second hypothesis examined the ability for crowdsourced workers to replicate the translational quality of professional translators. We took the professional translator scores as our gold standard, so we could examine the inter-annotator agreement (Cohen's Kappa) between the crowdsourced translations and the professional translations. We consider synonyms for a given term to be equivalent in our scoring. In Table 3, we group our results by multimedia genre instead of language, as this provides a more meaningful examination of their differences.

[5] www.worldlingo.com
[6] www.lingoes.cn
[7] www.pgups.ru/WebWN/wordnet.uix

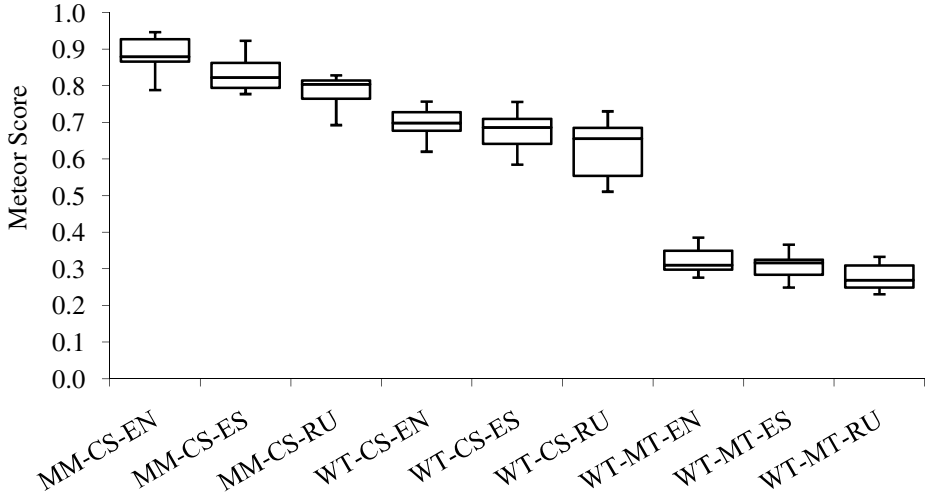


**Fig. 2.** Overview of Meteor scores comparing the results using all multimedia (MM) and the written transcripts only (WT). These are further grouped by translator type (MT or CS) and language (EN, ES and RU).

We are also able to observe from Figure 2 that there is a discernable difference between Meteor scores from crowdsourced translations (columns 4-6) and those from machine translation tools (columns 7-9), even when we only consider the written transcripts alone. When visual contextual clues are considered (i.e. compare columns 1-3 with columns 7-9 in Figure 2), we notice an even larger contrast, further validating our first hypothesis. The difference in the columns in Table 3 also show a gain in inter-annotator agreement when visual context clues are considered (recall that our gold standard translators had access to the multimedia versions of the video clips and written translations). We notice that the largest improvements are with the Music Videos (MV) – which had the most figurative language and therefore the most difficult for translation from a written document. This ratio of gains was consistent among all three languages.

**Table 3.** Inter-annotator agreement (Cohen's Kappa) between crowdsourced and professional translations grouped by genre. The MM column considers visual context clues whereas WT only considers the written transcripts.

| | **MM** | **WT** |
|---|---|---|
| TS | 0.69 | 0.61 |
| AN | 0.71 | 0.67 |
| MV | 0.65 | 0.57 |

Our primary interest is to examine differences in our four features: translation type, language, multimedia genre and use of multimedia or written transcripts only;

therefore we also represent our results as heat maps, which convey the differences between many of our features nicely. In Figure 3, we illustrate the difference in scores between the written transcripts for crowdsourced translations and online machine translations in a three-dimensional heat map.

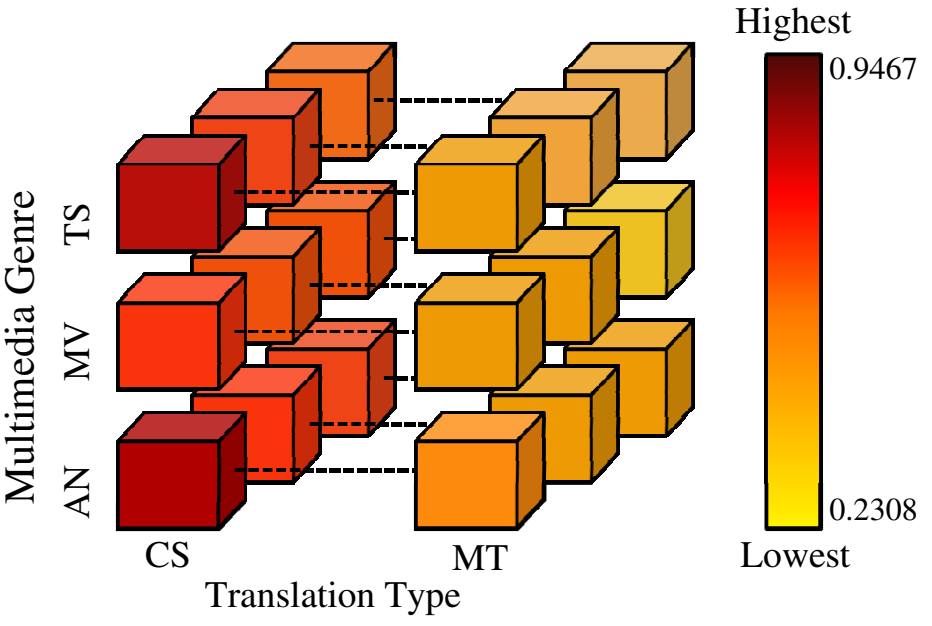


**Fig. 3.** Meteor scores for translations using visual context clues

From Figure 3, we can observe that the Meteor scores obtained from the crowdsourced translations and machine translations obtained from the written transcripts only are clearly different, and genre and language matter far less than translation type. Figurative language used in these three videos was made clearer through the use of visual context clues in the multimedia content. We also observe that the machine translations have far less variance across genre or language (the colors in the heat map are more uniform) compared with crowdsourced translations.

Figure 4 illustrates the difference in Meteor scores obtained using the visual contextual clues versus using only the written transcripts. The difference in scores ranged from +0.056 to +0.098, demonstrating the benefits of visual context clues as we have discussed earlier. However, with this heat map, we can see that English crowdsourced translations seem to benefit most, particularly those with the animated comedy skits (we believe this is likely due to the use of several situations in those videos involving a "play on words"). The difference in the use of visual context clues was smallest for Russian for the same genre (animated comedy skits). Initially we were not sure if this was due to both the multimedia and the written transcripts accounting for language that involved a play on words, or whether these comedic language constructs were missed entirely in the Russian translations; further examination of the Russian translations indicated the latter.

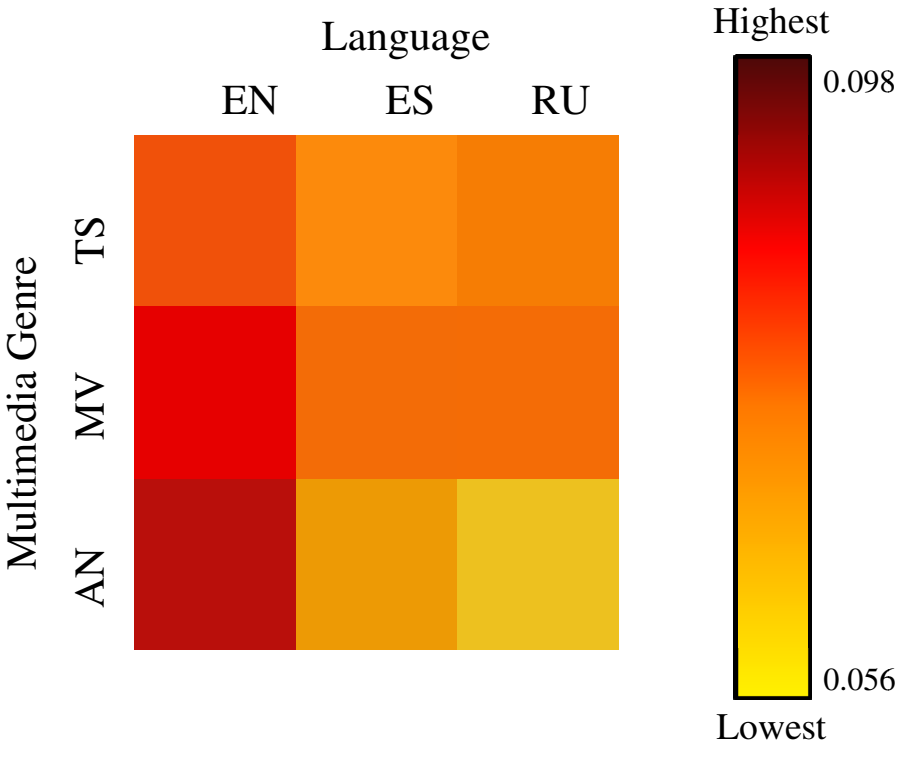


**Fig. 4.** Meteor scores for translations using visual context clues

Overall, the music videos appear to benefit most from the visual contextual clues. From casual inspection, the three music videos made substantial use of figurative language, which could be clarified through the use of visual aids in the multimedia content. As a result, we believe visual context clues enhance translation ability.

Some differences may be attributable to the Meteor scoring, not to translation ability. For example, "卑微地在人群中隐藏" in one set of lyrics was translated to Russian as "Осмелюсь скрываться в толпе", meaning "I dare to hide in the crowd". In Spanish, this same phrase is directly translated as "Humildemente escondidos en la multitude," the same as its direct translation in English: "Humbly hiding in the crowd." Although the difference in meaning is subtle, the Meteor score difference from the gold standard in the respective language is substantial, pointing out a weakness with Meteor, not with the translation itself. Therefore, to validate translations at a high level, we had human translators provide preference judgments on each feature:

- crowdsourced translations generated from written transcripts compared with crowdsourced translations generated from multimedia
- machine translations compared with crowdsourced translations
- professional translations compared with the crowdsourced translations

These blind preference judgments were able to validate the order of our earlier-reported Meteor rankings on each feature.

One additional issue we wished to validate was the ratio of costs between crowdsourcing translations and professional translations. Our professional translations were done on a per-word basis of 6-12 cents per word, for an average cost of US$49.65 per translation, and took an average of 4.7 business days to complete. Crowdsourcing translations were completed at an average cost of US$2.15 per translation (or 1/23$^{rd}$ of the cost of a professional translation) and took an average of 40 hours (1.6 days) to complete. Even if we require two translations to be done for each document to address any potential quality issues that may arise, we are able to still achieve a substantial savings over professional translations with only a small reduction in quality.

## 6 Conclusion

Our purpose in this paper was to determine other alternatives to machine translations that are low cost but more effective. We have investigated the role of visual contextual clues in multimedia translations, and the benefits they provide over translations from written transcripts. Additionally, we explored crowdsourcing's ability to provide the fast, cheap, and effective translations described in other studies. We examined these claims through the translation of nine Chinese language videos into three languages. We reported the results through heat maps, which are able to visually represent the relative differences between features.

The results of our study support our two hypotheses. Through a paired t-test, we verified that the visual context clues in our videos were able to increase in the Meteor evaluation scores at a p=0.05 level of significance over translations from written transcripts alone. In addition, we were able to achieve quality translations through crowdsourcing at a fraction of the cost of professional translations, demonstrated by strong inter-annotator agreement scores.

This study represents an initial foray into this area of translation. In the future, we plan to expand our study to include a larger set of videos and more genre variety and examine the role of these translations across a wider variety of languages. This will allow us to determine which languages rely more on visual context clues. In addition, we plan to measure how the translation quality differs between languages from a closely-related family versus languages from more distant families, across different genres.

## References

[1] Rao, L.: comScore: YouTube Reaches All-Time High of 14.6 Billion Videos Viewed In (May), http://techcrunch.com/2010/06/24/comscore-youtube-reaches-all-time-high-of-14-6-billion-videos-viewed-in-may/ (retrieved May 5, 2011)

[2] Crocker, M.: Computational Psycholinguistics. Kluwer Academic Publishing, Dordrecht (1996)

[3] Grainger, J., Dijkstra, T. (eds.): Visual word recognition: Models and experiments. Computational psycholinguistics: AI and connectionist models of human language processing. Taylor & Francis, London (1996)
[4] Johnson-Laird, P.N.: Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge University Press, Cambridge (1983)
[5] Chun, M.M.: Contextual cueing of visual attention. Trends in Cognitive Sciences 4, 170–178 (2000)
[6] Torres-Oviedo, G., Bastian, A.J.: Seeing is believing: effects of visual contextual cues on learning and transfer of locomotor adaptation. Neuroscience 30, 17015–17022 (2010)
[7] Deubel, H., et al. (eds.): Attention, information processing and eye movement control. Reading as a perceptual process. Elsevier, Oxford (2000)
[8] Mueller, G.: Visual contextual cues and listening comprehension: An experiment. Modern Language Journal 64, 335–340 (1980)
[9] Meskill, C.: Listening skills development through multimedia. Journal of Educational Multimedia and Hypermedia 5, 179–201 (1996)
[10] Fernald, A., et al. (eds.): Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. Developmental Psycholonguistics: On-line methods in children's language processing. John Benjamins, Amsterdam (2008)
[11] Roy, D., Mukherjee, N.: Towards Situated Speech Understanding: Visual Context Priming of Language Models. Computer Speech and Language 19, 227–248 (2005)
[12] Hardison, D.: Visual and auditory input in second-language speech processing. Language Teaching 43, 84–95 (2010)
[13] Cunillera, T., et al.: Speech segmentation is facilitated by visual cues. Quarterly Journal of Experimental Psychology 63, 260–274 (2010)
[14] Long, D.R.: Second language listening comprehension: A schema-theoretic perspective. Modern Language Journal 73 (Spring 1989)
[15] Gullberg, M., et al.: Adult Language Learning After Minimal Exposure to an Unknown Natural Language. Language Learning 60, 5–24 (2010)
[16] Kawahara, J.: Auditory-visual contextual cuing effect. Percept. Psychophys 69, 1399–1408 (2007)
[17] Lew, M.S., et al.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 2, 1–19 (2006)
[18] Zhang, X., et al.: A visualized communication system using cross-media semantic association. Presented at the 17th International Conference on Advances in Multimedia Modeling - Volume Part II, Taipei, Taiwan (2011)
[19] Tung, L.L., Quaddus, M.A.: Cultural differences explaining the differences in results in GSS: implications for the next decade. Decis. Support Syst. 33, 177–199 (2002)
[20] Morita, D., Ishida, T.: Collaborative translation by monolinguals with machine translators. Presented at the 14th International Conference on Intelligent User Interfaces, Sanibel Island, Florida, USA (2009)
[21] Bar-Hillel, Y.: A demonstration of the nonfeasibility of fully automatic high quality machine translation. Jerusalem Academic Press, Jerusalem (1964)
[22] Madsen, M.: The Limits of Machine Translation, Masters in Information Technology and Cognition, Scandanavian Studies and Linguistics. University of Copenhagen, Copenhagen (2009)
[23] Howe, J.: The Rise of Crowdsourcing. Wired (June 2006)

[24] Munro, R., et al.: Crowdsourcing and language studies: the new generation of linguistic data. Presented at the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010), pp. 122–130 (2010)
[25] Snow, R., et al.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. Presented at the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii (2008)
[26] Marge, M., et al.: Using the Amazon Mechanical Turk for transcription of spoken language. In: ICASSP (2010)
[27] Novotney, S., Callison-Burch, C.: Cheap, fast and good enough: automatic speech recognition with non-expert transcription. Presented at Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 2010), pp. 207–215 (2010)
[28] Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Presented at the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan (2005)
[29] Porter, M.: Snowball: A language for stemming algorithms (2001), http://snowball.tartarus.org/texts/
[30] Miller, G., Fellbaum, C.: WordNet, http://wordnet.princeton.edu (retrieved April 6, 2011)
[31] van Rijsbergen, C.: Information Retrieval, 2nd edn. Butterworths, London (1979)
[32] Agarwal, A., Lavie, A.: METEOR, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output. Presented at the Third Workshop on Statistical Machine Translation, Columbus, Ohio (2008)