Katherine Everitt & Christopher Lim & Oren Etzioni & Jonathan Pool & Susan Colowick & Stephen Soderland

Evaluating Lemmatic Communication

1 Introduction

The Internet has greatly expanded the ability to share information, enabling communication between physically and culturally distant people. However, there are over 6000 living languages (Lewis 2009), and the need to translate makes communication expensive even when distance is no longer an obstacle. Attempts to make translation inexpensive by automating it have been only partially successful, and they have ignored 99% of the world's languages. For example, the popular Google Translate application covers only 52 languages (Google 2010).

If people communicated using only lemmata (words and phrases in their citation, or dictionary, forms), automatic translation would be greatly simplified, permitting translation among thousands of languages. By combining existing resources (bilingual and multilingual dictionaries, thesauri, and glossaries), one could build a system that infers translations of arbitrary lemmata into arbitrary target languages.

In this paper, we evaluate such a system of translingual lemmatic communication. Senders encode message sentences into sequences of lemmata, these are automatically translated, and receivers attempt to decode the translated sequences of lemmata into sentences that reflect the meanings intended by the senders.

Lemmatic communication is a subtype of grammatically simplified communication. Morphological inflection is prevented, the use of grammatical particles is ineffective and deprecated, syntax is reduced to word order, and punctuation is unavailable. One can think of it as taking place in a "naturalistic controlled language" with unusually severe constraints (Pool 2006). Grammatically simplified communication has other subtypes in which the constraints on grammar are spontaneous rather than imposed. Among them are pidgin languages (Roberge 2009), foreigner talk (Ferguson 1975), and motherese (Whyatt 1994).

Lemmatic communication is also a case of the purely lexicographic (or "word for word") automation of discourse translation, an idea explored 50 years ago at Harvard University and the University of Washington (Hutchins 1986). At that time the goal was the unidirectional bilingual translation of literature, so it was deemed necessary to create a full-form lexicon in one language, or lemmatize inflected and derived forms in that language. In its new incarnation here, the goal is bidirectional panlingual translation of interactive communication. Compared with the original paradigm, this goal is both harder and easier to achieve. It requires translation across millions of language pairs, not only one. But it involves authors as participating encoders, so the system can translate lemmata rather than all lexical forms, and can thus derive the necessary data from already available bilingual and multilingual lexical resources. Interactive communication also provides additional opportunities for disambiguation, should confusion occur.

2 Lemmatic communication

The lemmatic communication process consists of three steps, as illustrated in Figure 1: encoding by the sender, automatic translation, and decoding by the receiver.

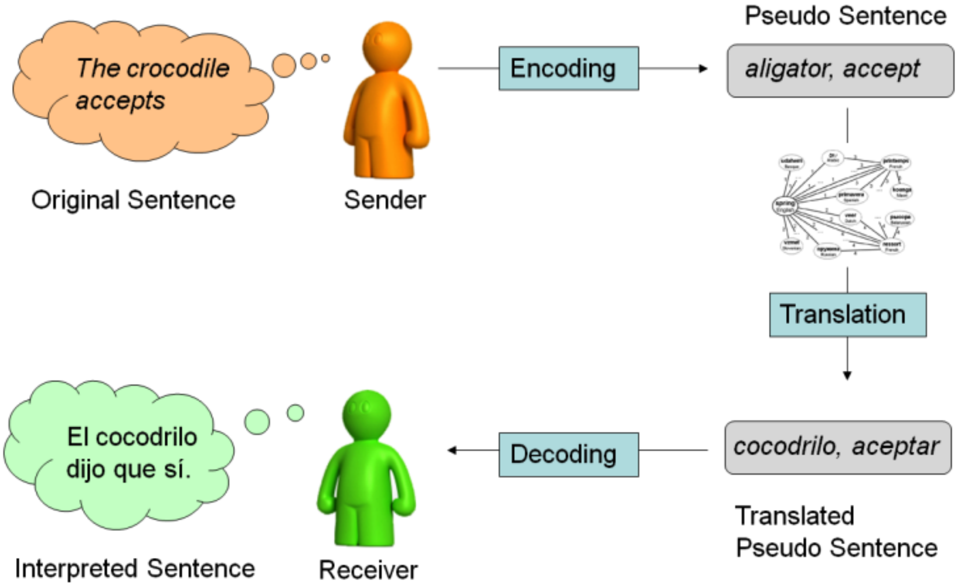


Figure 1: Lemmatic communication example

2.1 Encoding

In the encoding process, the sender selects lemmata and assembles them into a sequence, or pseudosentence, to convey a statement or question. For example, the sender might encode "A couple of previous guests recommended your hotel to us." as "two, previous, guest, recommend, hotel". Using autocomplete lists, the system permits the sender to select only lemmata that can be translated automatically into all of the foreseen target languages. Where it is known that there is only a single target language, the list will include all lemmata translatable into it, thereby offering a relatively large repertoire of source lemmata.

2.2 Translation

Translation is performed by TransGraph (Etzioni et al. 2007), a graph-based translation engine constructed from machine-readable lexical resources. The graph is under development, but it has currently made use of about 600 resources to discover about 12 million expressions in 1300 languages; 10 million senses represented with arbitrary numeric codes; and 27 million edges, each edge connecting a lemma to a sense. In the translation process, the system translates each lemma into a lemma in the target language and assembles the translated lemmata into sequences corresponding to the original sequences.

When one or more direct translations in the target language exist, the system translates the lemma into one of those. Otherwise, the system infers a translation from paths through intermediate translations. In each case, the system estimates the probability that each candidate translation is correct and selects the candidate with the greatest probability.

2.3 Decoding

In the decoding process, the receiver reads sequences of lemmata and attempts to infer the intended meaning of each sequence. For example, the receiver might read "my, home, inside, three, sleep, room, exist" and infer that the intended meaning is "There are three bedrooms in my home."

3 User study

We evaluated a system of translingual lemmatic communication to determine whether such communication can succeed and, if so, what conditions promote success.

Our questions were:

- How satisfying is the process of lemmatic communication to its participants?
- How long do encoding and decoding take?
- How much of the intended meanings is conveyed in lemmatic communication?
- Does the order of lemmata in a sequence convey useful information to the decoder?
- Is lemmatic communication less successful when the lemmata are translated than when they remain in the original language?

We performed the study on communications between speakers of Hungarian and speakers of Spanish. We chose these languages because they are typologically distinct, they were well represented in TransGraph at the time of the study, and we could get a significant number of speakers of them as participants.

We recruited participants through colleagues in Hungary and in Spanish-speaking countries. Our participants ranged in age from 20 to 72, with an average age of 44. They spoke on average 3.4 languages.

3.1 Encoding phase

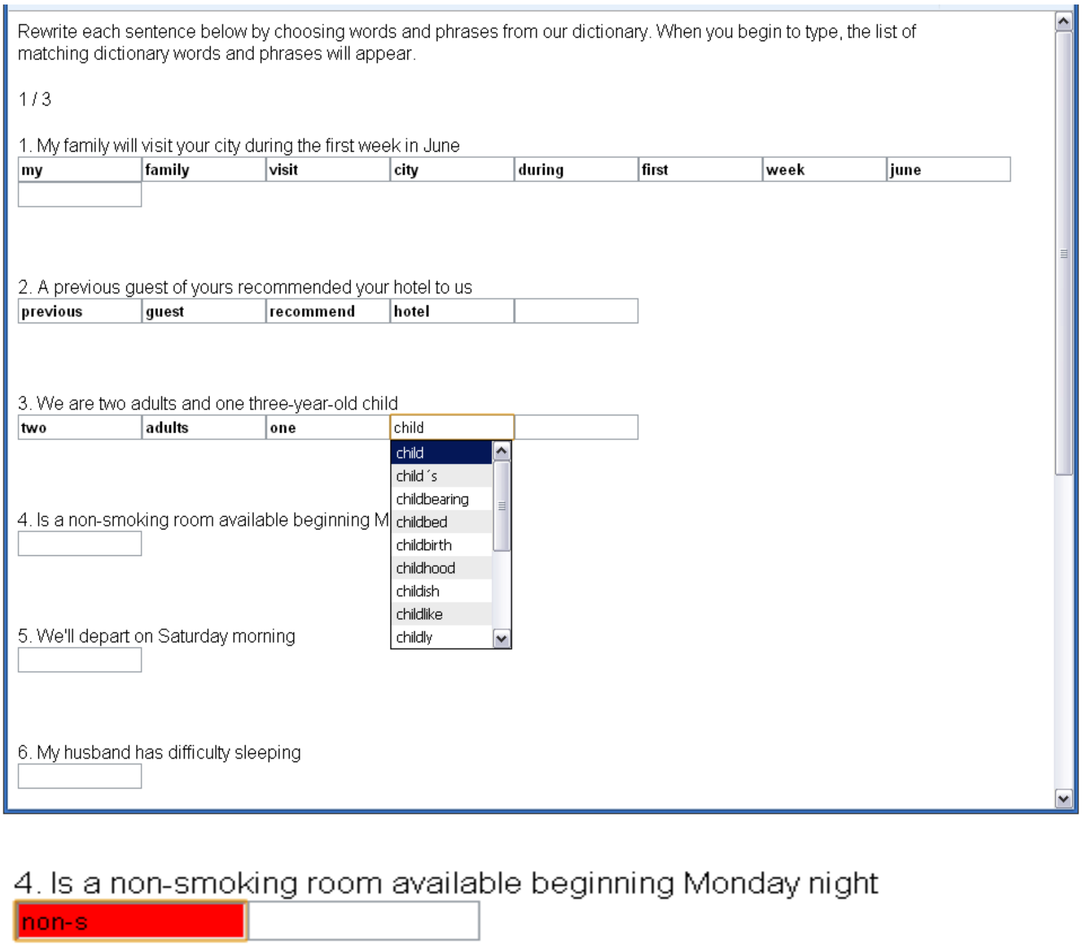
We created a set of three scenarios, each described with a series of ten sentences. The sentences were written in English, professionally translated from English into Spanish and Hungarian, and checked by bilingual translators.

Subjects in the encoding phase converted each sentence into a sequence of lemmata using our online encoding system. In addition to gathering encoded sequences for a later phase of the study, the purposes of the encoding phase were to get qualitative feedback from the encoders about the process and to get information (length, specificity, etc.) about likely encodings.

We used three scenarios:

- Visit: Visiting a city and booking a hotel
- Fable: The Monkey and the Crocodile
- Book Group: Message about a book group

Figure 2 shows the online encoding interface. It was written using .NET aspx pages and the jQuery JavaScript library (jQuery 2010) for dropdown functionality, with data stored in Microsoft SQL Server 2005. When encoders type two letters, a dropdown box appears showing the permitted lemmata. There were 18,139 permitted lemmata in Spanish and 24,482 in Hungarian. These were the lemmata in Spanish that TransGraph could translate into Hungarian and vice versa. If the encoder typed an incorrect string, the box would turn red and disallow it, as seen at the bottom of Figure 2.



Rewrite each sentence below by choosing words and phrases from our dictionary. When you begin to type, the list of matching dictionary words and phrases will appear.

1 / 3

1. My family will visit your city during the first week in June

my | family | visit | city | during | first | week | june

2. A previous guest of yours recommended your hotel to us

previous | guest | recommend | hotel

3. We are two adults and one three-year-old child

two | adults | one | child

4. Is a non-smoking room available beginning M

5. We'll depart on Saturday morning

6. My husband has difficulty sleeping

4. Is a non-smoking room available beginning Monday night

non-s

Figure 2: The online encoding interface

There were two Hungarian and two Spanish encoders. Encoders took between 6 and 50 minutes to encode each page, with a mean time of 17 minutes. The mean encoding time for a Spanish page was 9 minutes, versus 24 minutes for a Hungarian page. Because of the small sample size, encoding time may be skewed by one participant. However, the encoding times for Hungarian were always longer than for Spanish.

The mean encoded sequence length (in lemmata) was 1.17 more than the mean original sentence length (in words). Of all the sequences created, 68% were longer than, 17% were equally long as, and 15% were shorter than their source sentences.

3.1.1 Encoding feedback

The instructions given were deliberately imprecise, in order to explore people's natural inclinations. The instruction was "Rewrite each sentence below by choosing words and phrases from our dictionary." We also gave example encodings. From participant comments, we learned that participants often felt they needed to encode every word of the sentence. We also found that they wanted a way to encode information that is not available in the list, such as exclamations, questions and verb tense. In our pilot tests, which were conducted in English, there was excellent coverage of lemmata (40,957), and so participants expressed surprise when a specific lemma they wanted to use was not in the list. Also, participants were not very aware of the phrasal lemmata and occasionally had to go back and change multiple words into a corresponding phrase.

Encoders expressed some frustration with our list-constrained approach. One criticism was that a space does not move to the next box, so a tab or click is necessary. This is a necessary feature because phrases require a space to be typed without moving to the next box. Another criticism was that we required people to immediately correct their mistakes.

3.1.2 Encoding guidelines

We present a series of encoding guidelines based on our encoders' experience.

There is a tradeoff between allowing users to type any words they want in a traditional text-editor format and using a list-constrained approach. While the list constraint limits spontaneity, allowing people to type anything may cause a frustrating system response demanding changes to the lemmata that cannot be translated. Potentially, the most appropriate long-term solution is a combination of the two, where people are allowed to type what they wish but receive immediate feedback, such as a colored line under untranslatable lemmata with accompanying suggestions of alternatives. Another useful addition would be to automatically detect and combine phrasal lemmata or give clearer hints about their existence.

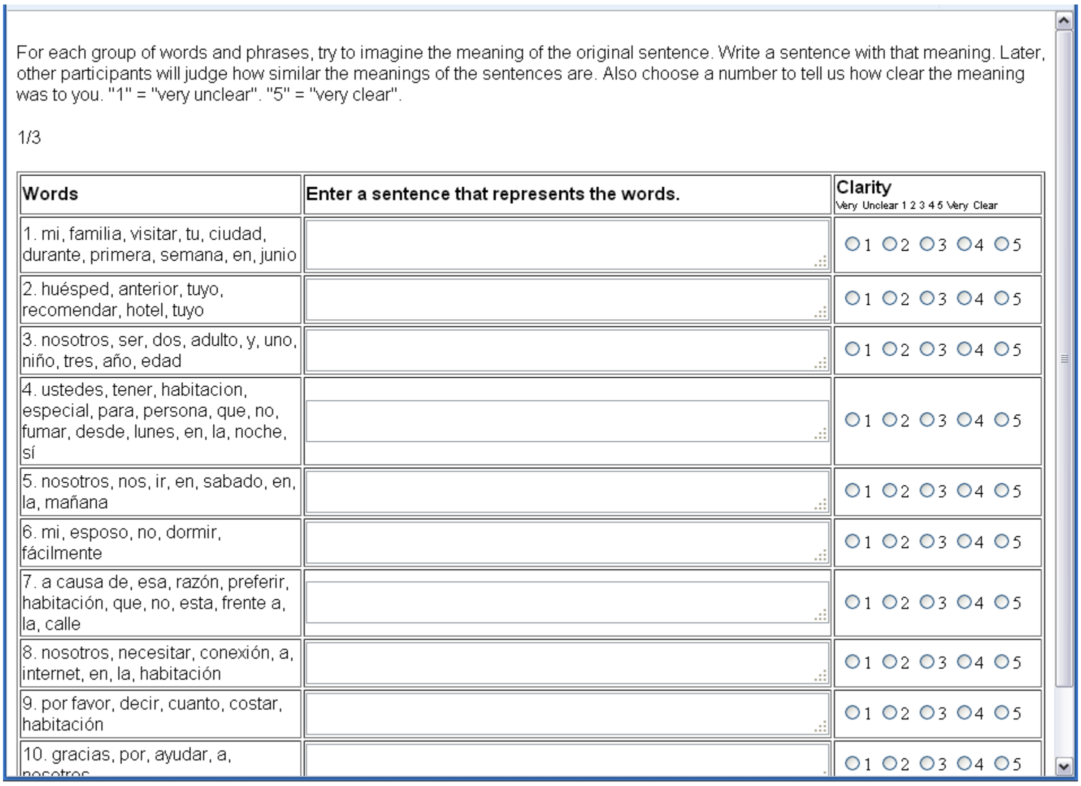
Some participant comments implied that the system would be more satisfying if it allowed the encoding of metadata describing properties of lemmata (e.g., tense) and of sequences (e.g., exclamation, question). Observation also suggests that it would be beneficial to encourage shorter encodings and to let encoders know that they do not need to encode grammatical particles and argument-marking words.

3.2 Decoding phase

The purposes of the decoding phase were to get qualitative opinions about the clarity of the lemma sequences and to collect sentences produced by decoders for comparison with the originals. Decoding took place under three conditions: Same, Randomized, and Translated. In the Same condition, the decoder was presented with one of the original encodings. In the Randomized condition, the decoder saw an encoding with the lemmata randomly re-ordered. In the Translated condition, decoders worked on an encoding whose lemmata had been translated from the other language by TransGraph, without any change in the order of the lemmata.

There were 49 decoding participants: 30 Hungarian-speaking and 19 Spanish-speaking. We presented the three scenarios in order: Visit, Fable, and Book Group, and counterbalanced the conditions of Same, Randomized, and Translation. All of the ten sequences within a condition were shown in order.

Figure 3 shows the decoding interface. The instructions were originally in the decoder's language but have been translated into English in Figure 3. Decoders expressed their guesses about the sequences' meanings by entering sentences and marked each sequence's subjective clarity on a scale of 1 to 5, where 1 meant very unclear and 5 meant very clear.



For each group of words and phrases, try to imagine the meaning of the original sentence. Write a sentence with that meaning. Later, other participants will judge how similar the meanings of the sentences are. Also choose a number to tell us how clear the meaning was to you. "1" = "very unclear". "5" = "very clear".

1/3

Words	Enter a sentence that represents the words.	Clarity Very Unclear 1 2 3 4 5 Very Clear
1. mi, familia, visitar, tu, ciudad, durante, primera, semana, en, junio		○1 ○2 ○3 ○4 ○5
2. huésped, anterior, tuyo, recomendar, hotel, tuyo		○1 ○2 ○3 ○4 ○5
3. nosotros, ser, dos, adulto, y, uno, niño, tres, año, edad		○1 ○2 ○3 ○4 ○5
4. ustedes, tener, habitacion, especial, para, persona, que, no, fumar, desde, lunes, en, la, noche, sí		○1 ○2 ○3 ○4 ○5
5. nosotros, nos, ir, en, sabado, en, la, mañana		○1 ○2 ○3 ○4 ○5
6. mi, esposo, no, dormir, fácilmente		○1 ○2 ○3 ○4 ○5
7. a causa de, esa, razón, preferir, habitación, que, no, esta, frente a, la, calle		○1 ○2 ○3 ○4 ○5
8. nosotros, necesitar, conexión, a, internet, en, la, habitación		○1 ○2 ○3 ○4 ○5
9. por favor, decir, cuanto, costar, habitación		○1 ○2 ○3 ○4 ○5
10. gracias, por, ayudar, a, nosotros		○1 ○2 ○3 ○4 ○5

Figure 3: Decoding interface

3.2.1 Decoding results

Figure 4 shows the mean subjective clarity by condition. As one would expect, the Same condition had the highest mean score. We also discovered that translated sequences without randomization were significantly less clear than randomized sequences without translation. Each difference shown in Figure 4 was significant ($p<0.01$). The mean result for the Translated condition—the one in which lemmatic communication might actually be put to use—was 2.99.

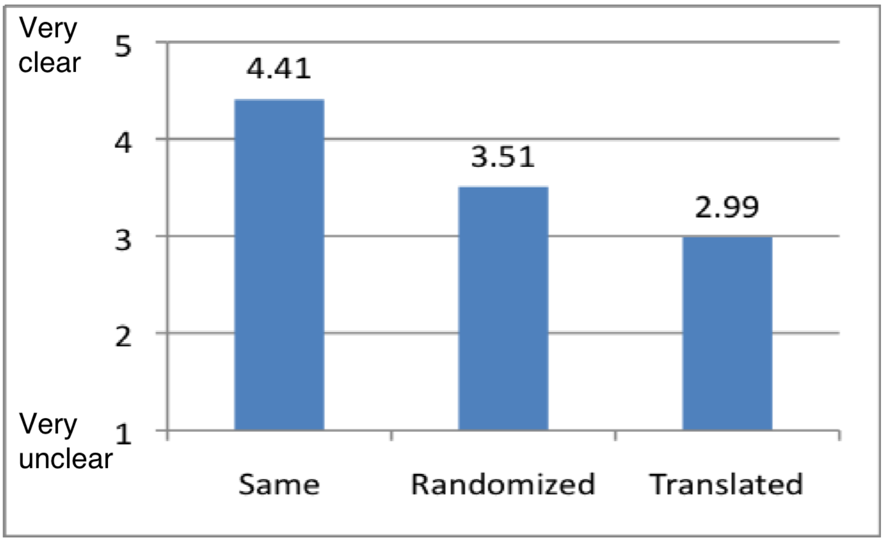


Figure 4: Subjective clarity by condition

Figure 5 shows the distribution of subjective clarity scores. For the Same condition, over half of the scores were 4 or 5. In the Randomized condition, over half were 3 or above. The Translate condition scores were fairly uniformly distributed. Almost 90% of sequences in the Same condition received a mean clarity score of 4 or 5, suggesting that lemmatically encoded messages can be understandable under the most favorable conditions. However, these proportions decreased to about 65% in the Randomized condition and 40% in the Translated condition.

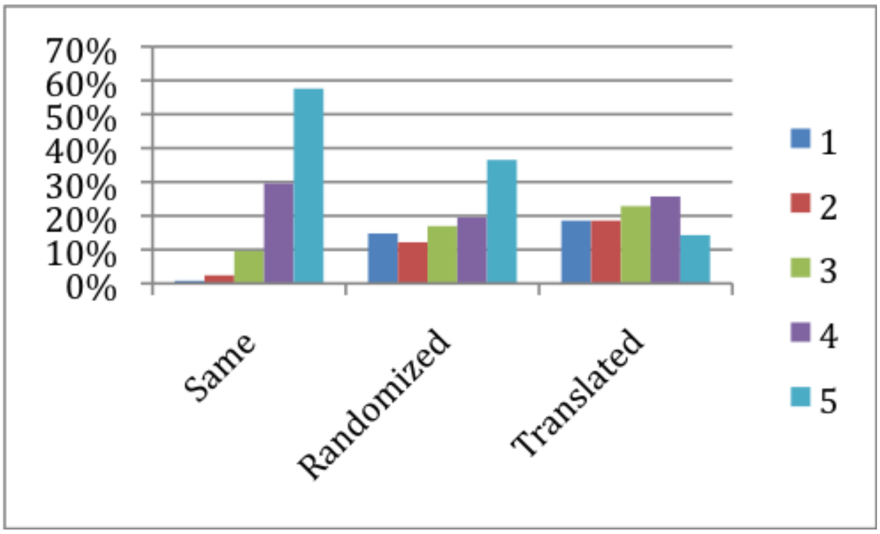


Figure 5: Subjective clarity distribution by condition

Figure 6 shows the subjective clarity by lemma sequence length. Longer sequences, over 11 words in length, had lower subjective clarity. Because longer sentences tend to have more clauses they are more susceptible to reordering effects (in the Randomized and Translate conditions) and mistranslation (in the Translate condition).

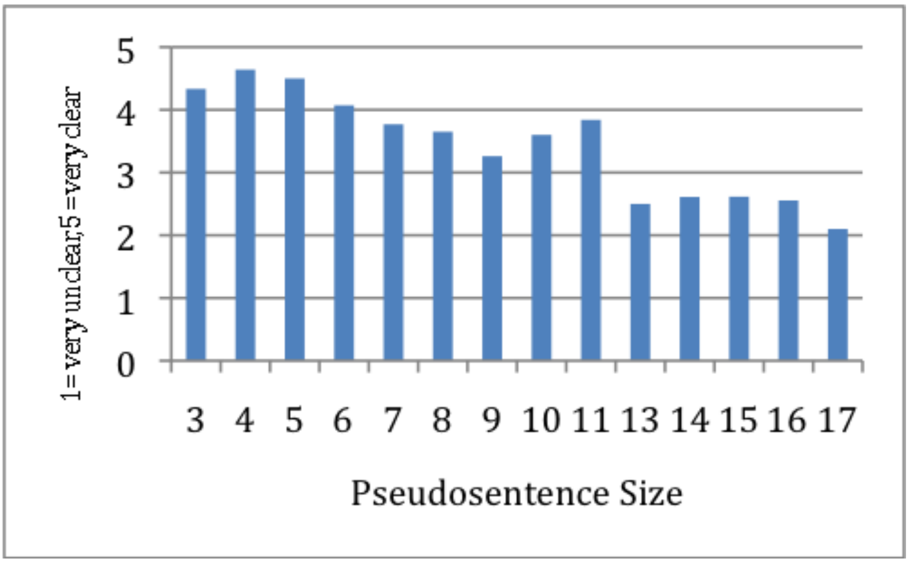


Figure 6. Subjective clarity by sequence length

Figure 7 shows the time to decode a sequence within each condition. On average, decoding took about a minute per sequence. The difference between the Same and Randomized conditions is marginally significant (p=0.067), and the difference between the Same and Translated conditions is significant (p < 0.01).

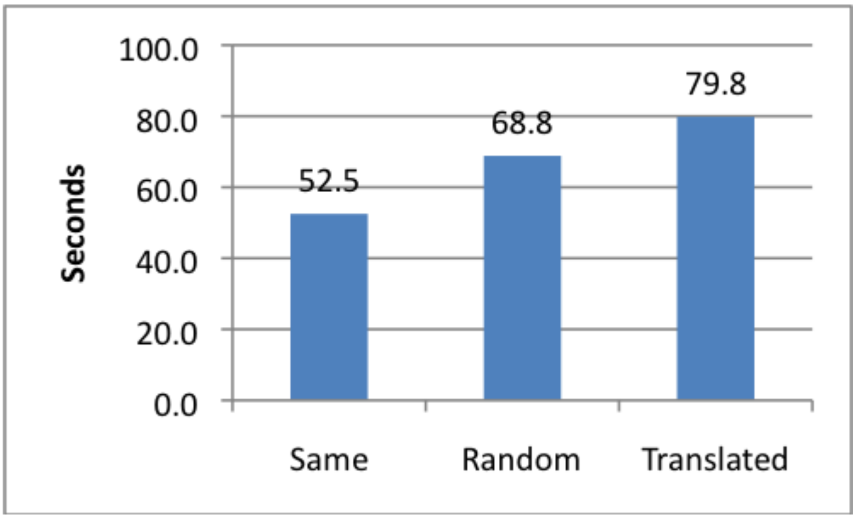
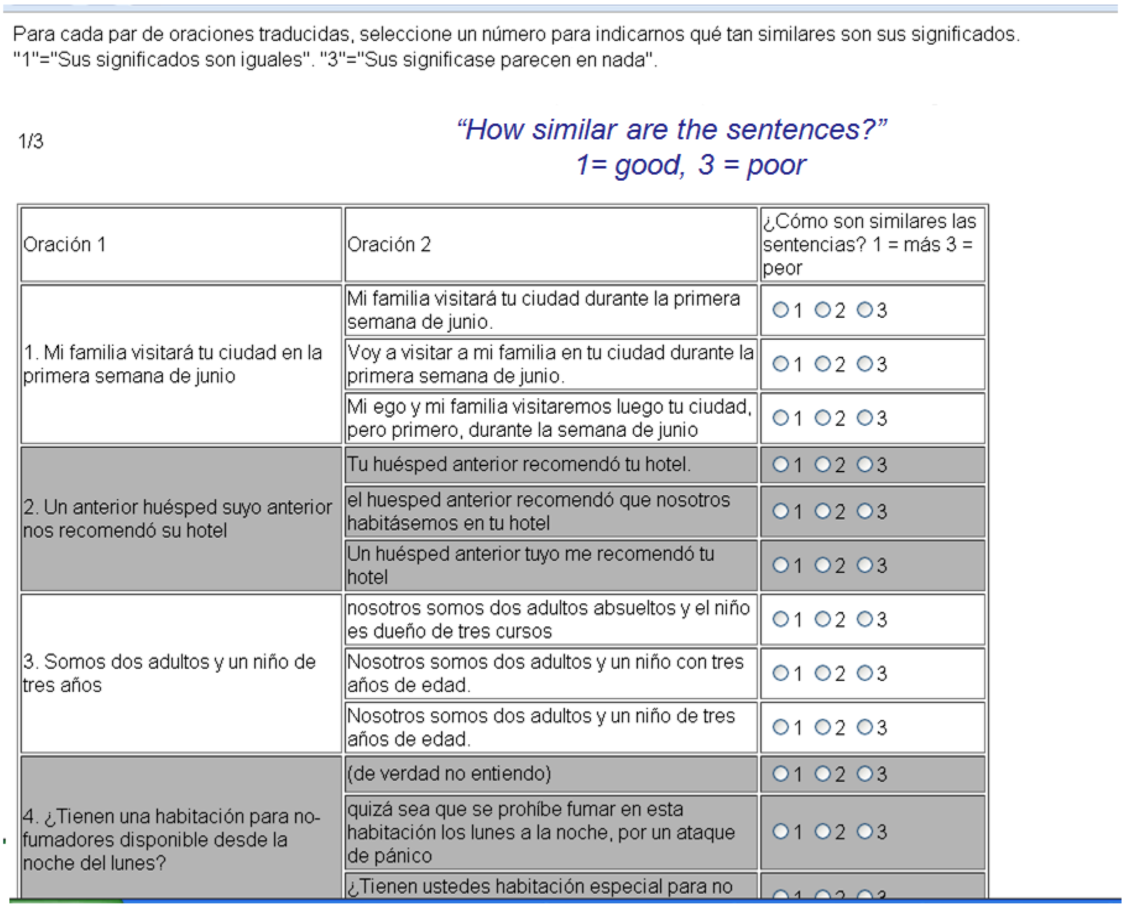


Figure 7: Mean time to decode a sequence within each condition

3.3 Evaluation phase

During the evaluation phase, the decoded sentences were compared with the original, professionally translated sentences and evaluated for meaning similarity. There were 10 Spanish-speaking participants and 12 Hungarian-speaking participants in this phase.

Figure 8 shows the evaluation interface. Participants were shown an original sentence and one Same, one Randomized, and one Translated version of that sentence, in random order. They were asked to score each output sentence in terms of the similarity of its meaning to the original sentence. A rank of 1 was good and 3 was poor.



Para cada par de oraciones traducidas, seleccione un número para indicarnos qué tan similares son sus significados.
"1"="Sus significados son iguales". "3"="Sus significase parecen en nada".

1/3

“How similar are the sentences?”
1= good, 3 = poor

Oración 1	Oración 2	¿Cómo son similares las sentencias? 1 = más 3 = peor
1. Mi familia visitará tu ciudad en la primera semana de junio	Mi familia visitará tu ciudad durante la primera semana de junio.	○1 ○2 ○3
	Voy a visitar a mi familia en tu ciudad durante la primera semana de junio.	○1 ○2 ○3
	Mi ego y mi familia visitaremos luego tu ciudad, pero primero, durante la semana de junio	○1 ○2 ○3
2. Un anterior huésped suyo anterior nos recomendó su hotel	Tu huésped anterior recomendó tu hotel.	○1 ○2 ○3
	el huesped anterior recomendó que nosotros habitásemos en tu hotel	○1 ○2 ○3
	Un huésped anterior tuyo me recomendó tu hotel	○1 ○2 ○3
3. Somos dos adultos y un niño de tres años	nosotros somos dos adultos absueltos y el niño es dueño de tres cursos	○1 ○2 ○3
	Nosotros somos dos adultos y un niño con tres años de edad.	○1 ○2 ○3
	Nosotros somos dos adultos y un niño de tres años de edad.	○1 ○2 ○3
4. ¿Tienen una habitación para no-fumadores disponible desde la noche del lunes?	(de verdad no entiendo)	○1 ○2 ○3
	quizá sea que se prohíbe fumar en esta habitación los lunes a la noche, por un ataque de pánico	○1 ○2 ○3
	¿Tienen ustedes habitación especial para no	○1 ○2 ○3

Figure 8. Evaluation interface

Figure 9 shows mean sentence similarity score by condition. As before, the best results occurred in the Same condition and the worst in the Translated condition. All differences were statistically significant ($p < 0.01$).

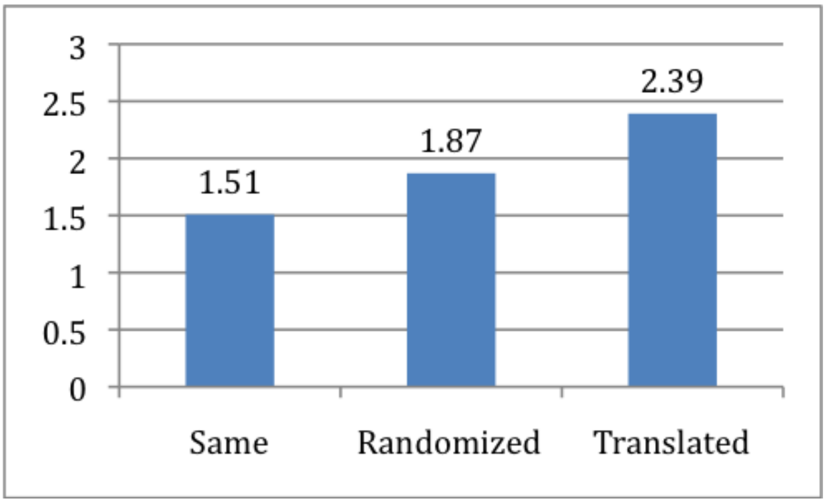


Figure 9: Mean sentence-similarity score by condition

Figure 10 shows the distribution of perceived sentence similarity by condition. It shows a much higher fraction of good responses in the Same condition than the Translated condition.

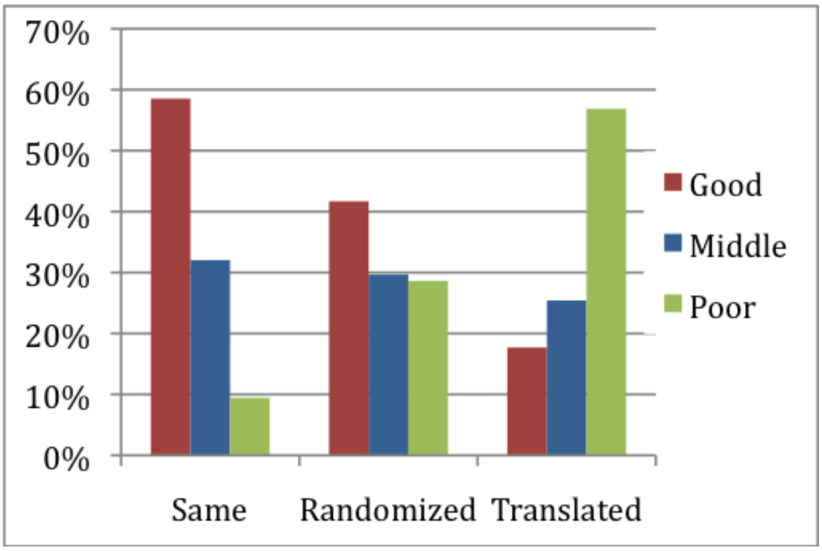


Figure 10: Perceived sentence-similarity distribution by condition

Figure 11 shows the distribution of perceived sentence similarity by scenario. The Visit and Book Group scenarios had a larger fraction of "good" responses, probably due to the better understanding of context surrounding them. The fable scenario was less familiar and had fewer "good" scores. The pairwise differences in mean score among the Visit, Fable, and Book Group conditions (1.87, 2.00, and 1.89) were all significant ($p < 0.01$).

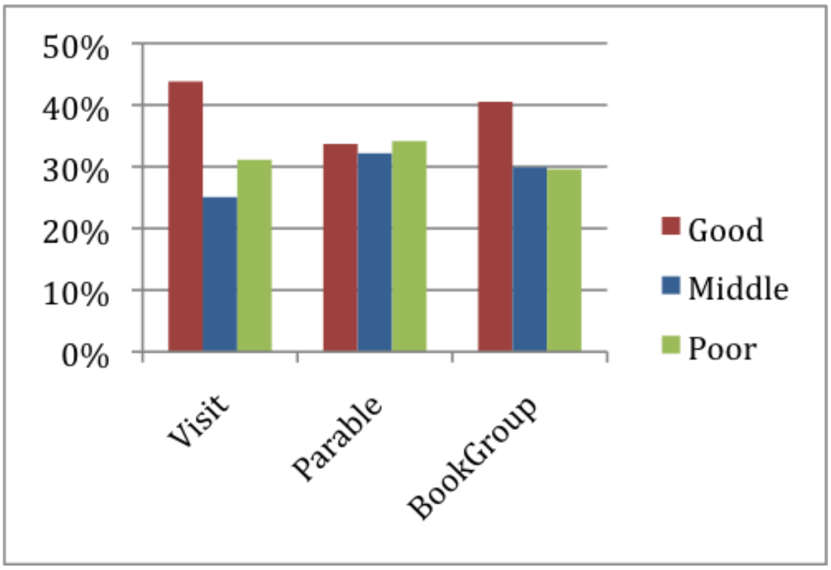


Figure 11: Perceived sentence-similarity distribution by scenario

Figure 12 shows the mean sentence length by perceived sentence similarity. Sentences marked Poor were significantly longer, on average, than those marked Good ($p < 0.01$) or Middle ($p < 0.01$).

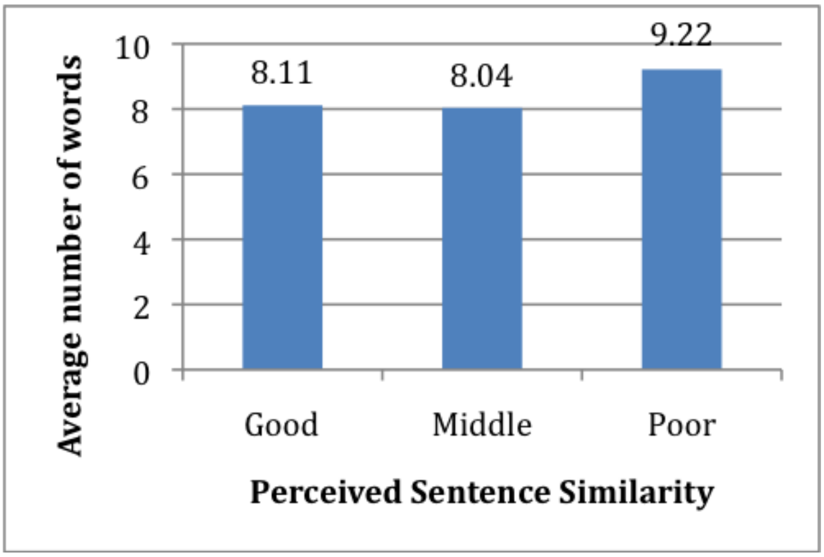


Figure 12: Mean sentence length by perceived sentence similarity

4 Discussion

Even with the confusion introduced by the randomization of lemma order and by translation, some successful communication between Hungarian and Spanish speakers occurred in our experiment. The Randomized condition sought to simulate (to an extreme degree) the independent effect of word-order differences among languages, separate from the effect of lemma translation. If this simulation is valid, we have evidence that word-order differences do impair lemmatic communication. Given the diversity of sentential (subject/verb/object) and phrasal (adjective/noun, etc.) word orders among languages and the intuitions that both encoders and decoders might develop to handle lemma ordering, much additional work could be done on the factors that detract from efficacy in lemmatic communication.

The efficacy of lemmatic communication may vary from language to language. Hungarian speakers took longer to encode lemmatic sequences than Spanish speakers, and it may be found that speakers of morphologically complex languages like Hungarian (Kiss 2002) find lemmatic communication more difficult, since it prevents the use of inflection to represent grammatical relations. Alternatively, second-grammar learning has been found partly independent of native-grammar features (Newmeyer 1983), and this may be true for those learning to encode and decode lemmatically, too.

The encoding system seems amenable to several improvements: making it faster, with fewer constraints on typing; permitting (and encouraging) encoders to split long sentences into multiple short lemma sequences; and permitting encoders to type freely and then get feedback on the translatability of what they have typed. Further efficacy could result from using more intelligent, context-aware translation; allowing the sender to check tentative translations (e.g., via back-translation feedback); or giving receivers access to multiple translation candidates. Converting our single-pass system to an interactive one, in which receivers can prompt senders for clarifications, might also permit the rapid resolution of linguistic uncertainty and mistranslations.

5 Conclusions

We have shown that lemmatic communication can work. Its translation component makes it considerably slower and more error-prone than other translation techniques, but it can be automatically extended to translate between the thousands of languages for which no pair-wise translation system exists. These results suggest that better interface design, the inclusion of annotation features, more intelligent translation inference, and sender-receiver interactivity could make lemmatic communication effective across thousands of languages.

6 Acknowledgments

We thank our participants, our translators, and this journal's anonymous reviewers.

7 References

Etzioni, Oren; Kobi Reiter, Stephen Soderland, Marcus Sammer (2007): "Lexical Translation with Application to Image Search on the Web." *Proceedings of Machine Translation Summit XI*: 175-182
– http://www.mt-archive.info/MTS-2007-Etzioni.pdf

Ferguson, Charles A. (1975): "Toward a Characterization of English Foreigner Talk." *Anthropological Linguistics* 17 [1]: 1-14

Google (2010): Google Translate
– http://translate.google.com/

Hutchins, William John (1986): *Machine Translation: Past, Present, Future*. (Ellis Horwood Series in Computers and their Applications.) New York: Halsted Press
– http://www.hutchinsweb.me.uk/PPF-TOC.htm

jQuery (2010): jQuery JavaScript Library
– http://jquery.com/

Kiss, Katalin É. (2002): *The Syntax of Hungarian*. Cambridge, U.K.: Cambridge University Press

Lewis, M. Paul, ed. (2009): *Ethnologue: Languages of the World,* Sixteenth edition. Dallas, Tex.: SIL International
– http://www.ethnologue.com/

Newmeyer, Frederick J. (1983): *Grammatical Theory*. Chicago: University of Chicago Press

Pool, Jonathan (2006): "Can Controlled Languages Scale to the Web?" *5th International Workshop on Controlled Language Applications*
– http://www.mt-archive.info/CLAW-2006-Pool.pdf.

Roberge, Paul T. (2009): "The Creation of Pidgins as a Possible Window on Language Evolution." Rudolf Botha, Henriëtte de Swart (ed.): *Language Evolution: The View from Restricted Linguistic Systems*. Utrecht: LOT, 2009
– http://lotos.library.uu.nl/publish/articles/000287/bookpart.pdf

Whyatt, Bogusława (1994): "Baby Talk—The Language Addressed to Language-Acquiring Children: A Review of the Problem." *Studia Anglica Posnaniensia* 29: 125-135
– http://ifa.amu.edu.pl/sap/files/29/10Whyatt.pdf

Authors

Katherine Everitt has a PhD in Computer Science from the University of Washington. Her research interests include human-computer interaction, collaborative interaction using tabletop surfaces and large displays, and applications for ubiquitous computing. She is currently working at Microsoft.
E-mail: everitt@cs.washington.edu
Website: http://www.cs.washington.edu/homes/everitt/

Christopher Lim has an MS in Computer Science & Engineering from the University of Washington, where he received a Microsoft Scholarship. His 2009 Master's Degree thesis is "Practical Translation for All Languages: The Design, Implementation and Evaluation of Panlingual Translator".
E-mail: chrislim@cs.washington.edu

Oren Etzioni is the Washington Research Foundation Entrepreneurship Professor of Computer Science & Engineering at the University of Washington. He is also the founder and director of the UW's Turing Center and a founder of several successful Internet enterprises.
E-mail: etzioni@cs.washington.edu
Website: http://www.cs.washington.edu/homes/etzioni/

Jonathan Pool is founder and president of Utilika Foundation, which supports research aimed at panlingual communication. He has been on the political science faculties of SUNY at Stony Brook and the University of Washington.
E-mail: pool@utilika.org

Website: http://utilika.org/info/pool.html

Susan Colowick is an information specialist and research associate at Utilika Foundation, where she has been involved in experimental studies, professional communications, and the acquisition of lexical data.
E-mail: smc@utilika.org

Stephen Soderland is a research scientist in the department of Computer Science & Engineering at the University of Washington. His recent research topics have been unsupervised information extraction from the Web and panlingual machine translation.
E-mail: soderlan@cs.washington.edu
Website: http://www.cs.washington.edu/homes/soderlan/