Monsieur, azonnal kövessen engem bitte![*] An Automatically Generated Interlanguage Tailored to Speakers of Minority but Culturally Influenced Languages

Luis A. Leiva and Vicent Alabau
ITI/DSIC, Universitat Politècnica de València
{luileito,valabau}@{iti,dsic}.upv.es

ABSTRACT

Automatic localization of cultural resources and UIs is crucial for the survival of minority languages, for which there are insufficient parallel corpora (or no corpus at all) to build machine translation systems. This paper proposes a new way to compensate for such resource-scarce languages, based on the fact that most languages share a common vocabulary. Concretely, our approach leverages a family of languages closely related to the speaker's native language to construct translations in a *coherent mix* of these languages. Experimental results indicate that these translations can be easily understood, being also a useful aid for users who are not proficient in foreign languages. Therefore this work significantly contributes to HCI in two ways: it establishes a language that can improve how applications communicate to their users, and it reports insights on the user acceptance towards the method.

Author Keywords
Localization, Machine Translation, Interlingua, Cityspeak

ACM Classification Keywords
H.1.2 User/Machine Systems: Human information processing; H.5.2 User Interfaces: Natural language; I.2.7 Natural Language Processing: Machine translation

INTRODUCTION

Language is one of the most notable (and, at the same time, vulnerable) treats of our cultural heritage. With an estimated number of roughly 7,000 languages spoken worldwide [7], it follows that the vast majority of languages are minority languages in every country in which they are spoken. As such, there exists a considerable amount of native speakers that may have trouble operating UIs, because of a non-familiar vocabulary, and accessing to cultural resources that are otherwise available in other (popular) languages. Moreover, it is commonly held that the diversity of languages should not be an obstacle to mutual understanding. This is where Machine Translation (MT) becomes a clear asset for humans (and computers) to communicate.

The traditional approach to MT is statistical, or data-driven, due to the scale at which it operates and the generalization capabilities it can provide. Under this paradigm, an MT system learns statistical parameters from a series of source-target language pairs. Therefore, a fundamental problem arises when no proper training data are available, as it often occurs with minority languages, as the MT system becomes useless. What is more, the number of potential bilingual translators of uncommon languages is hard to find, and so this is both expensive and not scalable. Fortunately, the expansive movements of migratory populations, combined with the historical dominance of some cultures, have resulted in a large common vocabulary among some linguistic areas, leading to what is known as *Sprachbunds* [4] — groups of languages that have become similar in some way because of geographical proximity and/or language contact. For instance, some examples of Sprachbund include the following:

Uralic: Estonian, Finnish, Hungarian, *Mordvinian*, *Kanthy*.
Nguni: Xhosa, Zulu, Ndebele, *Hlubi*, *Phuthi*.

In both cases, the first 3 languages are supported by some online translation systems. However, the other 2 are not. Overall, according to [7], it is estimated that more than a 10% of the world population could not be assisted by any MT system.

In this paper we exploit the Sprachbund concept to allow MT systems to handle minority languages. Our approach is entitled Culturally Influenced Interlanguage (CI^2), since we employ a family of languages closely related to the speaker's native language ($L1$) to construct translations in a *coherent mix* of these languages, i.e., preserving some features of $L1$. Note that it is assumed that the MT system has training pairs in those related languages, but the user is not proficient in any of them.

BACKGROUND AND RELATED WORK

The given context of our method is inspired by "Cityspeak", a polyglot street jargon devised in the film *Blade Runner* that incorporated words from many different languages (Spanish, French, Chinese, German, Hungarian, and Japanese) [8]. Needless to say this solution is not practical, as the users would need to learn a complex mix of highly unrelated languages. Historically, however, a large amount of effort has been invested in the development of new languages that would standardize a vocabulary common to the widest possible range of languages. Amongst these constructed languages, the most popular example is Interlingua, in which words may be taken from any language as long as they are present in a series of control languages. Interlingua tends

[*] Blade Runner's CitySpeak: *Sir, follow me immediately please!* [8]

CHI'12, May 5–10, 2012, Austin, Texas, USA.

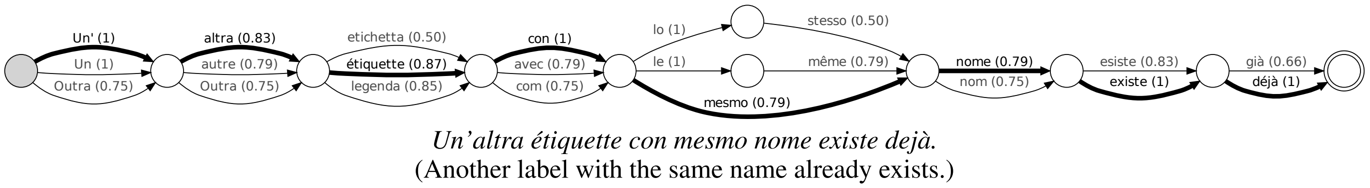


Figure 1: CI^2 can construct a sentence tailored to the Spanish language by using three related languages: Italian (top branches), French (central branches), and Portuguese (bottom branches). Each node in the graph is transited according to the maximum probability of belonging to the Spanish vocabulary (depicted in parentheses). As observed in the last 3 transitions, branches are collapsed either when a word is common to two (or more) of the base languages, or when no compatible alignments between such a base languages are found. Since this method is performed statistically, it can be completely extrapolated to other language families.

to be the lowest common denominator among such control languages (mostly Indo-European), providing also a simple grammar and regular word formation, but still it must be learned. Instead, a sentence could be formed in a way that it will be close to the vocabulary of $L1$. In these circumstances, it seems reasonable that people could easily understand multilingual translations, since such translations would be similar to their native language (Figure 1). In the same way as adaptive user interfaces try to fit the needs of a specific user (or group of users), CI^2 aims to construct a pseudo-language for the users so as to better communicate with them (Figure 2).

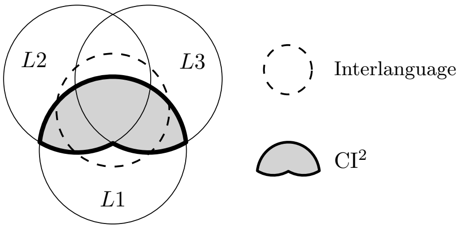


Figure 2: Artificially constructed interlanguages follow the Interlingua approach, i.e., they tend to the lowest common denominator among a series of control languages. CI^2, however, would employ only a subset of those control languages that intersect with the user's native language ($L1$).

To the best of our knowledge, there is little research that relates to our work; i.e., using multilingual information from different languages to better communicate the meaning of a message. For instance, Dagan et al. [2], to resolve lexical ambiguities in non-restricted text, concluded that "two languages are more informative than one". In Information Retrieval, Gollins and Sanderson [3] combined translations from two different languages to disambiguate search queries. The drawback of their approach is that it was limited to lemmatized dictionary entries, as presumably a common term would be present in the auxiliary languages involved. On the other hand, in HCI, the 'collaborative translation' concept has been explored. It is a fairly new approach to support monolingual translation (i.e., translation by people who speak only the source or target language, but not both) that unfortunately can be extremely slow (e.g., Hu et al. [5] reported an average speed of 40 sentences translated per day).

CI^2 SYSTEM OVERVIEW

Let us consider the following scenario: users whose native language is $L1$ want to understand a message in the source language S. Suppose that $L1$ is an under-resourced language, meaning that an automatic translation system cannot be built for it. At most, there is a vocabulary V available for $L1$, obtained e.g. from a regular dictionary. However, $L1$ belongs to a family of languages $L2, L3, \ldots, LN$ with a reasonable set of resources. Specifically, paired corpora are available to build a Statistical Machine Translation (SMT) system [6].

A first approach to overcome the problem is to build such an SMT system and hope that the user would be able to guess the meaning for any of the auxiliary languages. However, these translations may have vocabulary not understandable by the user. Our proposal is aimed to offer multilingual sentences in such a way that the communication channel maximizes the probability of understanding. CI^2 is defined as *a mix of languages tailored to the native language of the user*. This way, CI^2 tries to converge to a common ground by leveraging the resemblances that exist between the base languages that are used.

Formally, given a source string $\boldsymbol{s}$ and a vocabulary V for $L1$, the CI^2 system searches for the sentence $\boldsymbol{i} \in \mathrm{CI}^2$ with highest posterior probability, $\hat{\boldsymbol{i}} = \arg\max_{\boldsymbol{i}} \Pr(\boldsymbol{i}|\boldsymbol{s}, V)$. After some mathematical transformations we obtain the expression

$$\hat{\boldsymbol{i}} \approx \arg\max_{\boldsymbol{i}} \left[\Pr(\boldsymbol{i}|V, \hat{\boldsymbol{t}}_2, \ldots, \hat{\boldsymbol{t}}_N) \max_{\boldsymbol{t_2},\ldots,\boldsymbol{t_n}} \prod_{n=2}^{N} \Pr(\boldsymbol{t}_n|\boldsymbol{s}) \right] \quad (1)$$

where $\Pr(\boldsymbol{t}_n|\boldsymbol{s})$ are SMT systems built with Moses [6] for each auxiliary language. The first term is a graph that combines the best translations $\hat{\boldsymbol{t}}_2, \ldots, \hat{\boldsymbol{t}}_N$ and assigns to each word the probability of belonging to V. It is measured as the edit distance with respect to the closest word in the vocabulary, normalized by word length. Figure 1 illustrates this model.

EVALUATION

Due to the exploratory nature of this research, we carried out the experimentation in our department, which is located in Spain. Only for evaluation purposes, we assumed that Spanish was a minority language influenced by its neighboring countries: Portugal, Italy, and France. This way, we would have the necessary resources to perform the study: *1)* Spanish belongs to a known Sprachbund (Romance languages); *2)* there exists publicly available parallel corpora for this language family; and *3)* we can easily recruit a representative user sample of native speakers. The experiments consisted in translating Swedish (*Sv*) sentences to CI^2, which would be formed by a mix of Portuguese (*Pt*), French (*Fr*), and Italian

(*It*). We chose *Sv* as the source language so that participants would not understand the original sentences. The following hypotheses were evaluated:

- **H1**: CI^2 translations should look familiar to the user.
- **H2**: CI^2 translations should be easier to read and understand with respect to the base languages.

Participants

Seventeen volunteers (5 females) were recruited via email and word-of-mouth communication. The average age was 31.7 (SD=7.4). They were told to rank their level of language proficiency in a 1-5 Likert scale (1: not proficient, 5: almost native) for *Pt* [1.68 (0.8)], *Fr* [2.25 (1.0)], and *It* [1.87 (0.6)].

Design

To evaluate **H1**, we tested in Study 1 (S1) if there were differences regarding the number of unknown words among CI^2 translations and its 3 base languages. We carried out a one-way (continuous factor) analysis of variance (ANOVA), since we verified that both normality and homocedasticity did hold between groups. Interaction effects were considered at the $p < .05$ level for each of the tested conditions. On the other hand, **H2** was evaluated on the basis of the following criteria:

1. *Fluency*: Is the translation readable?
2. *Comprehension*: Is the translation understandable?
3. *Adequacy*: Regarding the reference sentence (in Spanish), is meaning really preserved?

In this second study (S2) we carried out two different experiments. In the first part we analyzed CI^2 translations in terms of the above mentioned criteria. In the second part we verified if any of the base languages would work better than CI^2.

Apparatus

We used the KDE4 corpus, which comprises the localization files of the popular software libraries, and is publicly available at the OPUS project [9]. This corpus has 92 paired languages and 8.89M sentence fragments. We trained an MT system with 167K sentences of each base languages (2.2M of running words), and reserved 100 *Sv* sentences for testing. Once the 100 test sentences were automatically translated to the base languages, we assembled the CI^2 translations according to Eq. (1). Next we selected two partitions for human evaluation. The first partition included 5 sentences, and the second one was formed by 10 sentences. All of them were randomly selected. We analyzed such translations and observed that the proportion of each language assigned by Eq. (1) was 25.2% *It*, 25.3% *Pt*, 12.6% *Fr*, and 36.9% common (including punctuation characters). Finally we developed two applications and a final survey to collect users' data.

Procedure

In S1, the sentences in the first partition were shuffled with their corresponding translations in *It*, *Fr*, and *Pt* (5 translations each). Overall, in this experiment users had to assess 20 sentences,which were presented in random order, by clicking on those words that were completely unknown to them. Table 1 summarizes the number of words for each language in both test partitions. ANOVA revealed that there were no statistically significant differences on word count $[F(3, 16) = 2.12, p = .137, \eta^2 = 0.33]$.

Study	# Sentences	Pt	Fr	It	CI^2
S1	5	16 (2.7)	18 (4.6)	16.6 (2.3)	12.8 (1.1)
S2	10	13.4 (2.2)	13.7 (2.6)	13.0 (3.1)	11.8 (2.7)

Table 1: Mean (and SD) number of words per sentence (per language) in both experiments.

Regarding S2, users were sequentially presented with a set of 10 *Sv* and CI^2 sentences. Each translation would be assessed in a 1-5 Likert scale based on the three criteria mentioned in Design section, followed by a ranking that would compare CI^2 and its base languages. To assess the adequacy criterion, users were given groundtruth Spanish translations of the source sentences. Eventually we acquired $17 \times 5 \times 4 = 340$ samples for S1, and $17 \times 10 \times 3 = 510$ samples for S2.

Finally, participants filled out a survey which was adapted from the System Usability Scale (SUS) questionnaire [1]. The goal was to measure the (subjective) user's appreciation towards the quality of CI^2 translations. Users could also complement the survey with freeform comments and ideas.

RESULTS

In S1, regarding the number of unknown words in each sentence, there were statistically significant differences between languages $[F(3, 327) = 18.25, p < .001, \eta^2 = 0.33]$. Post hoc comparisons using the Tukey HSD test indicated that the mean number of unknown words in CI^2 (Table 2) was significantly lower in comparison to *Fr* and *It* ($p < .001$ and $d > 0.45$ in both cases). However, for *Pt* there was no statistical significance regarding CI^2 ($p = .23, d = -0.36$). This result indicated that participants were confident while reading sentences both in CI^2 and *Pt*. Considering that no participant was proficient in that language, this can be explained because of the fact that Spanish and Portuguese have strong mutual influences.

Pt	Fr	It	CI^2
0.95 (0.9)	1.96 (1.5)	2.26 (1.2)	1.32 (1.0)

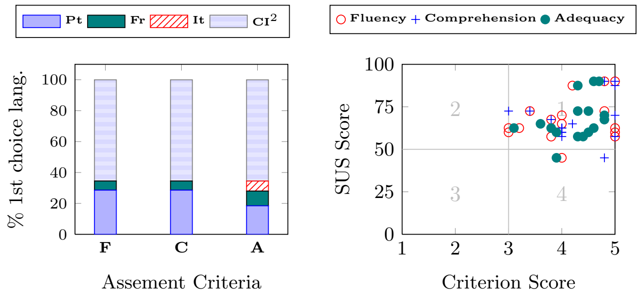
Table 2: S1 results. Mean (and SD) number of unknown words.

In the first part of S2, besides of the obtained high scores (Table 3), we did not find strong correlations between fluency, comprehension, and adequacy (in all cases, $0.1 < \rho < 0.3$). Nonetheless, in the second part of S2 it was interesting to observe that CI^2 translations were chosen in first place 64% of the time over the other languages for these cases (Figure 3a).

Fluency	Comprehension	Adequacy
4.14 (0.7)	4.26 (0.6)	4.26 (0.9)

Table 3: S2 results, part 1. Mean (and SD) criteria scores. Values were comprised between 1 (worse) and 5 (better).

Regarding the adapted SUS questionnaire, the average score was 67.65 (SD=12.16) — SUS scores are ranged between 0 and 100. This suggested that users were satisfied with CI^2 translations. We inspected each SUS question individually, and observed that, overall, they were scored as expected. For instance, *"I think that I could understand CI^2 frequently"* and *"I found CI^2 very interesting"* were notably ranked as positive; and, conversely, *"I found the sentences unnecessarily complex"* or *'I found CI^2 very cumbersome to understand"* were ranked as negative.



(a)

(b)

Figure 3: S2 results, part 2. [3a] Percentage of times a language was chosen in first position in terms of fluency (F), comprehension (C), and adequacy (A). [3b] SUS against these (averaged) assessment criteria. Quad. #1: Users that liked CI^2 and indeed it worked for them. Quad. #3: Users that did not like CI^2 and it did not work for them.

The viewpoint that users agreed most was that *"CI^2 is interesting and useful"*. One skeptical user reported that *"It surprised me! Translations were really helpful"*. Interestingly, some users remarked that *"CI^2 translations were really easy to deal with, sometimes ever better than the reference sentences"*. That is because the KDE4 corpus was automatically generated by an MT system following partial post-edition; therefore some sentences may not be completely error-free. Finally, although a few people did not find CI^2 very appealing (SUS < 50), Figure 3b shows that it worked quite successfully for them (score ≥ 4).

DISCUSSION

In light of the results, we envision a wealth of scenarios where CI^2 could be used. For instance, in some geographical areas several languages and dialects often do meet; however a few of them are official, and therefore governmental support is limited for the rest. In these cases, CI^2 would be an economic alternative to convey meaning, foster dialog between ethnic groups, and favor social integration. In this regard, we also believe that CI^2 may encourage people to improve their language skills, by making connections with the proposed vocabulary. Although it seems clear that users would not learn new grammatical structures with CI^2, we do hypothesize that it can be useful to learn new languages, or at least to be accustomed to them. For instance, the user could infer the meaning of new words by using the context of the sentences.

Importance to HCI and Limitations

Beyond the utility that our method provides from a general point of view, it can be a valuable asset for localizing software programs, as we have shown. Additionally, further applications in HCI exploring this potential include:

- Providing access to digital resources (e.g., web pages or electronic books) that cannot be translated by current systems to the minority languages.
- Bridging the gap between the linguistic competence of the user and the UI of the underlying computer system (e.g., selecting the appropriate vocabulary in command menus).
- Promoting the acceptance of UIs into developing countries that are still gaining access to technologies.
- Since CI^2 sentences are closer to the user's native language and hence are better understood, CI^2 can speed up the monolingual translation workflow in UIs like [5].

Finally, a strong limitation of CI^2 is that it would not work with related languages with unrelated alphabet glyphs (e.g., Romance languages have inherited many terms from Greek, however Greek uses a completely different alphabet set).

CONCLUSION

We have introduced CI^2, an automatically generated interlanguage that is tailored to the user (or groups of users thereof). It has been shown that CI^2 can serve as a communication channel across language barriers. Our research has been backed up with empirical evidence for Spanish speakers with little knowledge of foreign languages. CI^2 may also work with other language families, due to their mutual intelligibility. Future work includes replicating the same evaluation with other Sprachbunds, so that we can consolidate the foundations of our method. We hope that CI^2 will allow more people to share their knowledge with others worldwide, no matter which language do users or applications speak.

ACKNOWLEDGMENTS

This work is partially supported by MIPRCV (CSD2007-00018), erudito.com (TSI-020110-2009-439) and iTrans2 (TIN2009-14511) projects, and by GV/2010/067 and UPV/2009/2851 grants.

REFERENCES

1. Brooke, J. SUS: A "quick and dirty" usability scale. In *Usability Evaluation in Industry*. Taylor and Francis, 1996.
2. Dagan, I., Itai, A., and Schwall, U. Two languages are more informative than one. In *Proc. ACL* (1991), 130–137.
3. Gollins, T., and Sanderson, M. Improving cross language retrieval with triangulated translation. In *Proc. SIGIR* (2001).
4. Hickey, R., Ed. *The Handbook of Language Contact*. Wiley-Blackwell, 2010.
5. Hu, C., Bederson, B. B., Resnik, P., and Kronrod, Y. Monotrans2: a new human computation system to support monolingual translation. In *Proc. CHI* (2011), 1133–1136.
6. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. Moses: open source toolkit for statistical machine translation. In *Proc. ACL* (2007), 177–180.
7. Lewis, M. P., Ed. *Ethnologue: Languages of the World*, 16th ed. SIL International, 2009.
8. Sammon, P. *Future Noir: The Making of Blade Runner*. Orion Books, 1997.
9. Tiedemann, J. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Proc. RANLP* (2009), 237–248.