Tracking Inconsistencies in Parallel Multilingual Documents

Amit Pariyar, Donghui Lin, Toru Ishida
Department of Social Informatics
Kyoto University
Kyoto 606-8501 Japan
Email: amit@ai.soc.i.kyoto-u.ac.jp,lindh@i.kyoto-u.ac.jp,ishida@i.kyoto-u.ac.jp

***Abstract*—Translation practices in parallel multilingual documents are bound to generate inconsistencies in the documents due to missing information or part of document not translated, changes not propagated across languages, unavailability of same information in all languages. In this research, we present a mechanism to manage inconsistencies in parallel multilingual documents. To model inconsistencies in the multilingual contents, we propose a state transition model to define the states of sentences in the documents, the actions performed on the sentences and the set of transition functions. We then define consistency rules to signal the states of sentences resulting in inconsistencies. We illustrated the proposed mechanism with a case study on tracking inconsistencies in the multilingual MediaWiki Installation Guide as a tool support for tracking inconsistent portions such as missing information, changes not propagated and unavailability of same information in multilingual documents.**

***Keywords*-Multilingual document; Inconsistencies; State transition**

I. Introduction

Multilingual Document production is an important activity in the organization with the need to produce large amount of product documentations such as technical manuals, software documentations in several languages. With the support of tools and processes, organizations adopt various translation practices to produce multilingual documents. Pertaining to the evolving nature of the multilingual documents, change propagation and content reuse across multiple languages is not integrated into these translation practices.

The tendency against modifying the source document once the translation has started in sequential translations; the lack of synergy and content reuse across various languages in parallel authoring; the assumption of single master language usually English with all changes made in English before being translated into other language in incremental just in time translations are highlighted in [1][2]. More importantly, in multilingual context the inconsistencies in the documents such as missing information or part of document not translated, changes not propagated across languages, unavailability of same information in all languages cannot be ignored. With inconsistencies information cannot be accessed in native language though the information is available in other language. Managing inconsistencies in multilingual documents is therefore an important aspect in the translation practices. The capability of managing inconsistencies is also useful in knowledge management systems for accommodating and promoting knowledge resource consistently across multiple languages [3].

The practices in managing multilingual documents based on centralized representation of multilingual correspondences between the contents cannot deal in situations when the content in the multilingual documents are updated independently in native languages by the authors [4] . In such situations, the translation of contents in producing multilingual documents also requires to determine the source content in a particular language that is to be derived in other languages [5]. However inconsistencies in multilingual contents due to missing parts, change propagation, unavailability of same information adds further to managing multilingual documents.

In supporting consistent change propagation, the multilingual authoring tools in [2] [6] relies on expertise in making changes to conceptual models e.g. knowledge base. The multilingual community participation as in Microsoft MSDN wiki [7] for generating multilingual contents require a tool support for notifying updated contents, missing contents in different languages for enabling content sharing. To ensure the consistency of contents in multiple languages and to support the collaboration of the multilingual communities, a mechanism is needed to coordinate collaboration activities with a tool support for tracking inconsistent portions in the multilingual documents.

In this paper, we present a mechanism to manage inconsistencies in Parallel Multilingual Documents. To model inconsistencies in multilingual contents, we propose a state transition model to define the states of the parallel aligned sentences in the multilingual documents, set of actions performed on these sentences and the set of transition functions that describe the state transition of the sentences. To check for inconsistencies, we define set of consistency rules to signal states of the sentences leading to inconsistencies. We then illustrate the proposed mechanism in tracking inconsistencies for missing information or part of document not translated, change propagation and unavailability of same information in multilingual MediaWiki Installation Guide.

This paper is organized as follows. In Section 2, we present a motivating example to illustrate inconsistencies

978-0-7695-5047-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CultureComputing.2013.11

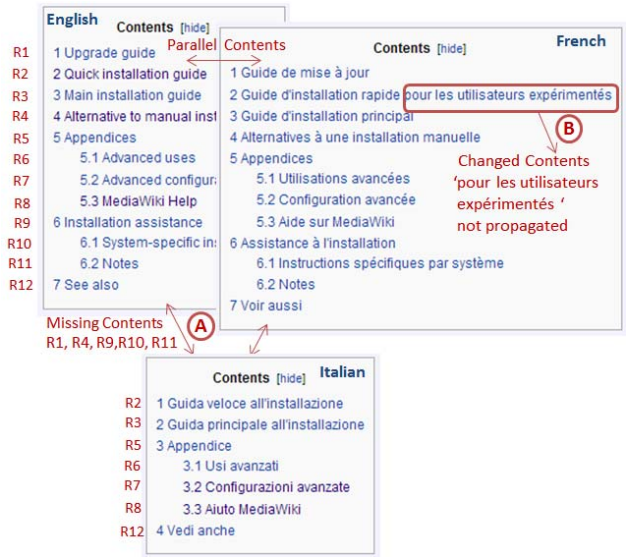


Figure 1. Inconsistencies of contents in MediaWiki Installation guide

in multilingual documents. The mechanism for managing inconsistencies in parallel multilingual documents is presented in Section 3. A case study illustrating the proposed mechanism as a tool support for tracking inconsistencies in the multilingual MediaWiki Installation Guide is presented in Section 4. The related works are discussed in Section 5. Finally we make a conclusion in Section 6.

II. Motivating Example

To illustrate our motivation for this research, we select as an example: support manuals for MediaWiki[1] as it is available in multiple languages, with many revisions to the contents, exhibits inconsistencies across languages and managed by the collaborative efforts of the communities. We inspect the technical manuals for Installation guide[2] on MediaWiki which is available in 22 languages.

*A. Missing information or part of document not translated:*It is evident from the content box in each of the language versions of the article, that the contents are not available equally in all languages. However it is observed that the English and French versions of the installation guide are parallel aligned documents corresponding same contents which is obvious in the case of technical manuals. The Italian version of the installation guide does not have the complete information. Fig.1 shows the parallel contents $\langle R1, R2, ..., R12\rangle$ between the English and French versions with some missing contents $\langle R1, R4, R9, R10, R11\rangle$ in the Italian version.

B. Lack of propagation of changes: In order to check for the propagation of changes across the languages, we modified contents in $\langle R2\rangle$ to the Installation guide in French language by adding 'pour les utilisateurs expérimentés' meaning 'for experienced users' and checked for its availability in the aligned sentences in other languages or some kind of notification in other language version for indicating the presence of the availability of contents to be updated. We find that the feature for propagating changes in different language version of the Installation guide in such collaborative setting is not available.

Lack of synergy and content reuse: In understanding the variation of contents, we inspect the revision history of the contributions made to the articles in each of the language versions and find that the contributions are isolated and language specific clearly lacking synergy or communication between the communities. From the contribution history, it is clear that though the date of updates in Italian version of installation guide[3] is recent, the contributors are not aware of new contents that is already available in other languages. In the French version of the revision history, it is also found the contributor commenting as 'Updating according to the last English version' which means that translation activities are being performed to update contents from another language.

Similar characteristics can also be found in other support manuals for MediaWiki such as User help, FAQ with inconsistencies in multilingual contents. Our main concern in this research is to manage such inconsistencies and support collaboration activities in tracking inconsistent portions in the multilingual documents. With the presence of such tool, the multilingual community can collaborate in generating multilingual contents and keep track of the updated contents or missing contents and share the contents with each other. In the next section we discuss our mechanism in detail.

III. Mechanism for Managing Inconsistencies

In our work we extend the correspondences of segments in parallel multilingual documents presented in [5] with the information about states of the sentences. For content reuse, we support leveraging contents from monolingual document and generate parallel aligned sentences with the consistency of information maintained at the sentence level in the parallel multilingual documents.

The notation used throughout this paper is explained here. A monolingual document D_l is a document D with contents available in language l. A Parallel Multilingual Document PMD is a set of monolingual documents in several languages, all being the translation of the same source document. This means $PMD = \{D_e, D_f, D_i\}$ is the set of monolingual documents that represents source document in English D_e and the corresponding translations from the source document to produce Italian D_i and French D_f documents for maintaining consistent information in all languages. The parallel aligned sentences R_{ke}, R_{kf} and R_{ki}

[1]http://www.mediawiki.org/wiki/MediaWiki
[2]http://www.mediawiki.org/wiki/Manual:Installation_guide
[3]http://www.mediawiki.org/wiki/Manual:Installation_guide/it

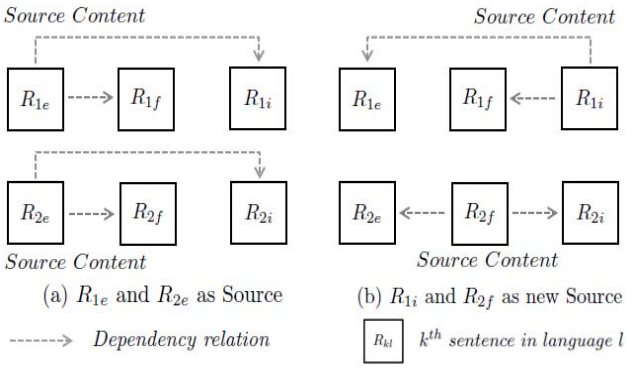


Figure 2. Dependency between parallel aligned sentences in PMD.

are the k^{th} aligned sentence in English, French and Italian documents.

We will first illustrate the state transition model to define the states, actions, state transition of the sentences during content modifications. We refer to [8] for the basic concepts in automata theory. We then define set of consistency rules to signal states of sentences leading to inconsistencies.

A. State Transition Model

The state transition model can be described as a tuple: $M = (Q, \Sigma, \delta, q_\circ, q_f)$ where

(1) $Q = \{S, T, I, M\}$ is the set of states of the sentences corresponding to Source, Translated, Intermediate and Modified states respectively. $q_\circ$ and q_f are initial and final states.

(2) $\Sigma = \{add, delete, update, mdf\_content, mdf\_quality, translate, \varepsilon\}$ is the set of actions performed on the sentence.

(3) δ is the state transition function, $\delta(q, x) = q'$, refers to the change in state from q to q' for some input action x where $q \in Q$, $q' \in Q$ and $x \in \Sigma$.

States. To explain states, consider Fig.2. which depicts the relation between the source of content and its translation between parallel aligned sentences in the multilingual documents. In Fig.2(a) the sentence $(R_{1e}, R_{2e}) \in D_e$ is the source of content and the sentences $(R_{1f}, R_{2f}) \in D_f$ and $(R_{1i}, R_{2i}) \in D_i$ holds the content derived from its source due to translation. With dependency relation, the translation of content is directional from the source sentence holding the latest information. In the multilingual context, as contents are modified in multiple languages, the sentence holding the latest updated information i.e. source sentence changes.

In Fig.2 (b) sentence $R_{2f} \in D_f$ and $R_{1i} \in D_i$ appear as new source sentence holding the latest information. The dependency is established by translating contents from R_{2f} to parallel aligned sentences R_{2i}, R_{2e} and from R_{1i} to parallel aligned sentences R_{1f}, R_{1e}. The sentences appear as source sentence or translated sentence as initial or final state in the document. We use this illustration to define the states for the sentences R_{kl} in the multilingual document.

Source. A sentence R_{kl} in the multilingual document is said to be in Source state (S) in language l, if it holds the latest information i.e. content and is the source for translation into other language.

Translated. A sentence R_{kl} in the multilingual document is said to be in Translated state (T) in language l , if it holds the translated content from the parallel aligned source sentence. In Fig.2(a). R_{le} and R_{2e} are Source state and the parallel aligned sentences R_{1f} and R_{2f} are Translated state.

Intermediate. A sentence R_{kl} in the multilingual document is said to be in Intermediate state (I) in language l , if operation such as add or delete, causing modification to content in sentence are not saved to become modified content.

Modified. A sentence R_{kl} in the multilingual document is said to be in Modified state (M) in language l , if the content in the sentence is modified but requires specifying modification purpose to determine as a new source of content for translation. For e.g. content modified by adding facts is a new source of content and requires to be propagated i.e. translated in other language; content modified for improving translation quality do not need propagation.

At a sentence level, during content modifications in the document (say English document D_e), the states of the sentences in the document D_e at the certain time snapshot τ is $State(D_e)^\tau = \langle S, T, I, S, ...M \rangle$.

Actions. The action ***add*** defines the addition of new information and **delete** defines the deletion of existing information from the sentence. The action ***update*** saves the modification made to the content. The action ***mdf_content*** and ***mdf_quality*** checks the (a) modification of content with the addition or deletion of information i.e. need translation (b) modification of the content's quality e.g. improving quality of translation i.e. do not need translation. The action ***translate*** translates the content of sentence from the source to the corresponding parallel aligned sentences. The ***automated action*** ε is triggered due to change in dependency relation as in Fig.2.(b) so that the aligned sentences corresponding to a new Source becomes Translated state.

Transition Function. The state transition of sentence in Table 1 is used to deduce the state of the sentence when actions are performed during modification.

B. Consistency checking

We define consistency rules to represent the states of sentences in multilingual documents that correspond to inconsistency during the modification of contents and direct the translation of multilingual contents from its source. We categorize the consistency rules into two groups.

Table I
STATE TRANSITION TABLE FOR A SENTENCE

State Q	Action Σ						
	add	delete	update	mdf_quality	mdf_content	translate	ε
Source (S)	I	I	-	-	-	S	T
Intermediate (I)	I	I	M	-	-	-	-
Modified (M)	I	I	M	S or T *	S	-	-
Translated (T)	I	I	-	-	-	-	T

** depending upon initial state, if $q_o = S$ then S else T*

Translation Consistency rules govern the translation of multilingual contents in multilingual documents by representing the state for which the content needs to be translated. Below are the rules for translation consistency.

1) Translation from Source state: The multilingual contents in Source state in the document is the source of content and has to be translated to corresponding parallel aligned sentence in other documents.

2) Translation from Source state due to content modification: The Source state reached from modifying the content (addition or deletion of facts or information) in the sentence is new source and is the candidate for translation in corresponding languages.

Transition Consistency rules govern the transition of states during the modification and represents states leading to inconsistency. Below are the rules to follow for checking inconsistencies.

3) Initial and final states: The initial and final state of sentence is either Source or Translated state so that it can either be the source of content or hold the contents translated from other parallel aligned sentence.

4) Presence of multiple source : For the parallel aligned sentences, $\langle R_{kl1}, R_{kl2}, R_{kl3}, ..., R_{kln} \rangle$ the sentence in Source state in multiple languages results in multiple sources of content leading to unavailability of same information in all languages as they are no longer translation of each other.

5) Dependency relation: As a sentence becomes a new source and the contents are translated from the source, the state of the corresponding parallel aligned sentences changes to Translated state if they are either in Source state or Translated state with an automated action ε to maintain the dependency relations.

6) Presence of Intermediate state: The content of sentence in Intermediate state leads to contents not updated to become available in other languages. The content of sentence has to be saved by the update action so that the modification is preserved in the sentence.

7) Presence of Modified state: The content of sentence in Modified state has to be translated to be made available to document in other languages so it is required to check the purpose of modifications. A sentence in modified state leads to missing information or part of document not translated into other languages.

8) Purpose of Modification determines a new source : In Modified state the modification to the sentence with mdf_content becomes new source and needs propagation while the sentence modified with mdf_quality in not a new source for translation. This allows the changes in content in the document in one language to be propagated in other languages.

Consistency Checking uses both translation and transition consistency rules to identify the states of sentences that violate these rules and signal for state of the sentence that results in missing information or part of the document not translated, unavailability of same information and ensures change propagation in all documents.

In the next section we will illustrate the proposed mechanism for tracking inconsistencies in generating multilingual contents.

IV. CASE STUDY: TRACKING INCONSISTENCIES IN MULTILINGUAL MEDIAWIKI INSTALLATION GUIDE

In this case study, we show the use of state transition model and consistency checking in tracking inconsistencies during collaboration in generating multilingual contents on the Installation guide for MediaWiki. The steps carried out in this case study are presented below:

Step 1: Leverage multilingual contents. To prepare for multilingual collaboration, with the major concern for generating same contents across multiple languages and the support for content reuse as required in the Installation guide for MediaWiki, we emphasize on leveraging contents from the source language as the initial step. Referring to Fig.1. the contributors in English community generate the contents for the Installation guide in English due to domain expertise in MediaWiki. The content in English is reused to generate the multilingual contents in French and Italian languages respectively. From state transition model, the states of the parallel aligned sentences $\langle R_1, R_2, R_3, ..., R_{31}, ..., R_{kl} \rangle$ in English document D_e are in Source state (S) and parallel aligned sentence in French D_f and Italian D_i documents are in Translated state (T).

Step 2: Modify contents in multilingual documents. We performed modification to the contents in English, French and Italian documents as presented in Fig.3. For the English document we improved the quality of content in $\langle R_{21} \rangle$ by changing the term 'copy' to its plural form 'copies'. We also updated contents to the existing sentence $\langle R_{22} \rangle$ and $\langle R_{23} \rangle$ by adding more information and added a new sentence in $\langle R_{24} \rangle$. We added more information to the existing sentence

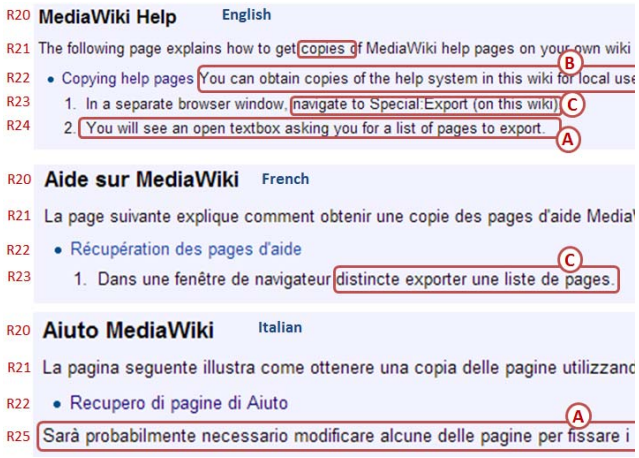
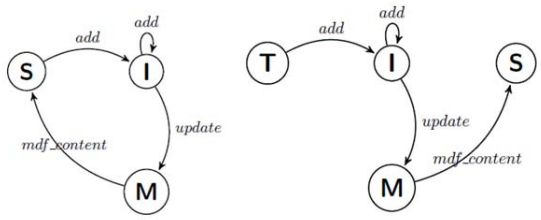


Figure 3. Inconsistencies from modifying contents in multilingual documents.



(a) modified R_{23} in English (b) modified R_{23} in French

Figure 4. State transition in parallel aligned sentences

$\langle R_{23} \rangle$ in French document and a new sentence $\langle R_{25} \rangle$ in Italian document. The state transition of the parallel aligned sentences in multilingual documents during modification is as in Fig.4. which shows for the case when content is modified to sentence $\langle R_{23} \rangle$ with initial state as Source in English and Translated in French documents. Our main focus is to support multilingual communities in tracking inconsistencies in multilingual contents during modification.

Step 3: Track Inconsistencies. To facilitate tracking inconsistencies, we integrate information about states of the parallel aligned sentences in the multilingual documents along with the violation of rules if it exists in a tabular representation as in Fig. 5. In this step, we check the following:

A. Missing Information part of document not translated: The missing of parallel sentences $\langle R_1, R_4, R_9, R_{10}, R_{11}, R_{23}, R_{24} \rangle$ in Italian, $\langle R_{24}, R_{25} \rangle$ in French and $\langle R_{25} \rangle$ in English documents depicts violation of rule (1) which requires contents to be translated from the source as it is available. E.g. The consistency rule (1) applied to sentence $\langle R_{25} \rangle$ in Italian document generates inconsistencies as part of document is not translated. Inconsistencies from the missing information or part of document not translated is tracked in Fig.5.

Aligned Sentences	States in Parallel Multilingual Documents			Rule Violation
	D_e	D_f	D_i	
R_1	S	T		(1)
R_2	S	S*	T	(2)
R_3	S	T	T	
R_4	S	T		(1)
...	...	...	...	
R_{11}	S	T		(1)
R_{12}	S	T	T	
...	...	...	...	
R_{20}	S	T	T	
R_{21}	S	T	T	
R_{22}	S*	T	T	(2)
R_{23}	S*	S*		(4)
R_{24}	S			(1)
R_{25}			S	(1)
...	...	...	...	
R_{30}	T	S	M	(7)
R_{31}	T	I	S	(6)
...	...	...	...	
R_{kl}	S	T	T	

S* : new Source

Figure 5. Tabular view for tracking Inconsistencies in parallel aligned sentences.

B. Change propagation across languages: To identify propagation of changes we refer to the changes made to the contents in step 2. With rule (8) the quality modification to the English content in $\langle R_{21} \rangle$ is not a new source of content and does not need propagation i.e. translation. But the contents to the existing sentence $\langle R_{22} \rangle$ in the English creates inconsistencies as the updated content is not available in French and Italian. The consistency rule (2) is applied to $\langle R_{22} \rangle$ in the English document to detect the need to propagate changes i.e. translation in other languages.

C. Unavailability of same information. For the parallel aligned sentence $\langle R_{23} \rangle$ the addition of more information both in English and French documents create inconsistencies as these contents are both new source and no longer translations of each other. The state transition of parallel aligned sentence $\langle R_{23} \rangle$ in French and English documents during modification resulting both as new source of content is depicted in Fig.4. The consistency rule (4) is applied to $\langle R_{23} \rangle$ to detect the presence of multiple source. Inconsistencies from the unavailability of same information from multiple source is tracked in Fig.5.

With the tabular view for representing the states of the parallel aligned sentences along with the rules violation as in Fig.5. we illustrated the support for tracking inconsistencies due to missing information or part of document not translated, change not propagated and unavailability of same information in multilingual Installation Guide. The support tool with features for tracking inconsistencies can be useful in managing inconsistencies in producing multilingual documents in a collaborative settings.

V. Related Works

In previous approaches, the focus of managing multilingual contents has been in maintaining multilingual corre-

spondences between documents [4][5] . Al-Assimi et al. presents a mechanism for collecting modified segments in multiple languages and producing coherent multilingual documents [5]. In our work we also emphasize on multilingual correspondences but for the situations when multilingual contents are updated in the native language by the authors and the inconsistencies are bound to encounter. The centralized representation of multilingual correspondences between the contents in [4] cannot deal in this distributed context. We also consider source of content to be important in propagating changes within the documents during translation. With the expertise needed for making the changes in the conceptual model for propagating changes consistently in multilingual authoring tools such as DRAFTER, PILLS [2][6] we focus on simplified task of making changes to contents directly with necessary support in tracking changes in different languages. The use of languages resources such as Language Grid already useful in multi-language discussion platform for translation activities [9][10] can also be used in such authoring activities. Some of the constraints in translation practices for content reuse is explored in LizzyWiki and WikiBABEL which provides insight into the usefulness of tool support in tracking inconsistencies in community collaboration [1][7]. However managing inconsistencies in document descriptions from analysis model, specifications in general do not consider multilingual aspect [11].

Compared to previous works, we focus on managing multilingual documents particularly in managing inconsistencies due to missing information, unavailability of same information, changes not propagated across language. In our approach, we emphasize on these cases causing inconsistencies in multilingual documents and also suggest a feature on tracking inconsistencies as tool support which is not available in community collaboration for generating multilingual contents.

VI. Conclusion

In this research we presented a mechanism to manage inconsistencies in parallel multilingual documents with a support for tracking inconsistent portions in the multilingual documents. For modeling inconsistencies in multilingual contents, first we proposed a state transition model to define the states of the parallel aligned sentences in the multilingual documents, actions performed and the set of transition functions. We then define set of consistency rules to check for states of the sentences resulting in inconsistencies due to unavailability of same information, missing information or part of document not translated and changes not propagated across languages.

From the case study we illustrated the proposed mechanism as a tool support for tracking inconsistent portions in multilingual MediaWiki Installation Guide. With the tabular view for highlighting inconsistencies in the parallel aligned sentences, we are able to track the missing information, changes not propagated across languages and unavailability of same information in multilingual documents. To test the practicability of the proposed mechanism we plan to implement a prototype tool for tracking inconsistencies and integrate it into multilingual document production environment such as wiki for producing software documentation collaboratively in multiple languages.

Acknowledgement

This research was partially supported by Service Science, Solutions and Foundation Integrated Research Program from JST RISTEX, and a Grant-in-Aid for Scientific Research (S) (24220002) from Japan Society for the Promotion of Science.

References

[1] A. Désilets, L. Gonzalez, S. Paquet, and M. Stojanovic, "Translation the wiki way," in *Int. Sym. Wikis*, 2006, pp. 19–32.

[2] A. Hartley, "Multilingual document production from support for translating to support for authoring," *Machine Translation*, vol. 12, pp. 1–2, 1996.

[3] D. E. O'Leary, "A multilingual knowledge management system: A case study of fao and waicent," *Decision Support Systems*, vol. 45, no. 3, pp. 641 – 661, 2008.

[4] P. Tonella, F. Ricca, E. Pianta, and C. Girardi, "Restructuring multilingual web sites," in *Proceedings of International Conference in Software Maintenance*, 2002, pp. 290–299.

[5] A. Assimi, A. Basset, and C. Boitet, "Management of non-centralized evolution of parallel multilingual documents." in *Proceedings of 10th International World Wide Web Conference*, 2001, pp. 19–32.

[6] N. Bouayad-Agha, R. Power, D. Scott, and A. Belz, "Pills: Multilingual generation of medical information documents with overlapping content," in *Proceedings of LREC*. MIT Press, 2002, pp. 2111–2114.

[7] A. Kumaran, K. Saravanan, and S. Maurice, "wikibabel: community creation of multilingual data," in *Proceedings of the 4th International Symposium on Wikis*. ACM, 2008.

[8] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation (3rd Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2006.

[9] A. Hautasaari and T. Ishida, "Analysis of discussion contributions in translated wikipedia articles," in *Proceedings of the 4th international conference on Intercultural Collaboration*. ACM, 2012.

[10] T. Ishida, Ed., *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*. Springer, 2011.

[11] B. Nuseibeh, S. Easterbrook, and A. Russo, "Making inconsistency respectable in software development," *Journal of Systems and Software*, vol. 58, pp. 171–180, 2001.