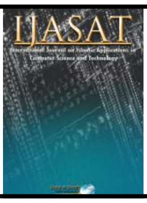


Preparation of an Islamic Parallel Bilingual Corpus for Deaf People with 3D Animations

Yahya O.M. Elhadj[1], Kamel Ayadi[2,3], Ahmed Ferchichi[3]

[1] Sabbatical Leave @ IRIT, University Paul Sabatier, France
[2] Community College, Prince Sattam bin Abdulaziz University, Saudi Arabia
[3] ISG, Université de Tunis, Tunis, Tunisie
[1] yelhadj@irit.fr, [2] ayadi_kamel1@yahoo.fr, [3] ahmed.ferchichi@isg.rnu.tn

Abstract

An Initial Islamic bilingual parallel corpus of Arabic and Sign language was developed in a previous research we carried out at Al-Imam University, Saudi Arabia. This corpus is composed of Arabic texts describing the basic Islamic topics such as prayer, pilgrimage, fasting, etc. in terms of elements and "functioning". In this paper, we present our work related to the refinement and enhancement of this corpus and its structure; we present also a prototype of a teaching and learning environment that we will continuously improve to reach a good educational support for deaf. This environment will be powered by a hybrid (rule-based and statistical approaches) translation component combined with avatar-based 3D animations, which are currently under development.

Keywords: Deaf, Sign Language, Corpora, Parallel Corpus, 3D animation, Avatar.

1. Introduction

Sign Language is the natural mean of communication between deaf, almost as talking is for healthy persons. To communicate with the community around them, deaf need an interpreter to translate signs to equivalent words and vice versa. Two major problems emerge: the lack of a permanent interpreter and the presence of a third person, which can be considered as violation of their intimacy. Thus, there is a pressing need to find appropriate means to facilitate social integration for deaf and make easy their communication with other communities around them. Many techniques were built to assist blind, deaf and people who are partially or totally paralyzed, as well as people who have difficulties with learning and understanding. In America and many European countries, these techniques helped millions of disabled to acquire their independence, reduce their need for other people, improve their education and release their potential. However, the Arab deaf are still less served. Recently, some research works appeared in the literature providing Arab deaf with tools and applications for both education and communication purposes, but they are still very limited; See for instance Mohandes (2006), (Jemni & Elghoul 2007), (Jemni & Elghoul 2008), Halawani (2008), Al Ameiri (2011).

In this context, we have initiated an ambitious project to help Arab deaf improving their access to educational resources (with a focus on the Islamic subject) by progressively, building a translation system from Arabic to Sign Language, creating 3D animations using Avatar technologies, and finally developing an appropriate learning environment gathering all needed components. The focus on Islamic topics was guided by three essential factors: 1) the fact that educational programs in the Arab countries are heavily based on religious backgrounds, 2) the centrality of learning and disseminating teachings of Islam to all segments of the society, 3) the potential for expansion of the work, knowing that the lexicon and keywords for Islamic subjects are not just common to Arab only, but it could easily be expanded to all Muslims around the world.

First stages of this project were devoted to the development of an initial Islamic bilingual parallel corpus composed of Arabic texts covering the basic religious fields and their translation into Sign Language to be used as infrastructure of the work (Elhadj et al. 2012a, 2012b, 2012c). In this paper, we present our work related to the refinement and enhancement of this corpus and its structure to be suitable for machine translation purposes.

2. Overview of the Initial Islamic Bilingual Parallel Corpus

Arabic Sign Language Resources, such as, linguistic structures, unified dictionaries, corpora, and so on, are, unfortunately, still almost absent in many Arabic countries. It is obvious that without these language resources the development of a well-founded ICT (Information and Communication Technology) applications serving the deaf community is not possible. To contribute in this regards, we initiated a work to develop a bilingual parallel corpus of Arabic and Sign Language in the Islamic domain[1] (Elhadj & Zemirli, 2014); Arabic texts describing the basic Islamic topics such as prayer, pilgrimage, fasting, etc. in terms of elements and "functioning" were firstly collected, prepared, and then translated to Sign Language into two ways: textual translation written by hearing interpreters, and visual translation directly recorded from deaf signers (see table 1 and figure 1 below for an extract). An alignment between these two translations was also performed (see figure 2 for an example).

Table 1: Example of the Textual Translation

ID	الجملة العربية(Arabic Sentence)	الترجمة الإشارية بالعربية (Signe Sentence - translated)
1	كتاب الصلاة	كتاب + صلاة
2	تعريف الصلاة	(تعريف + تعريف) + الصلاة
3	الصلاة لغة الدعاء	الصلاة + لغة + الدعاء
4	وشرعاً قربة فعلية ذات أقوال وأفعال مخصوصة	شرعًا + (قربة + قربة) + أقوال + أفعال + صلاة + ركوع + مخصوصه
5	حكم الصلاة	حكم + الصلاة
6	الصلاة فرض بالكتاب والسنة والاجماع	الصلاة + فرض + (الكتاب + الكتاب)+ (السنة + السنة) + (الاجماع + الاجماع + الاجماع)
7	فمن أنكر ذلك فهو مرتد عن دين الإسلام بلا خلاف	دين + اسلام + عن + فاضي + شيخ + علماء + اختلاف + لا
9	باب أوقات الصلاة	باب + صلاة + وقت
10	للصلاة أوقات محدّدة لا تؤدّى خارجها، وهي	حدود + وقت + صلاة + حد + اذان + قبل + لا + حد + اقامة + بعد + لا
11	صلاة الظهر	صلاة + ظهر
12	أول وقتها هو زوال الشمس	وقت + أول + صلاة + ظهر + شمس عاموديه
13	وآخر وقتها إذا صار ظل كل شيء مثله	صلاة + ظهر+ حركة الشمس + ظل شئ + مثل+ وقت + اخر
14	مضافا إليه القدر الذي زالت الشمس عليه	صلاة + ظهر + حركة الشمس + ظل + أضافة

This bilingual parallel corpus can be used to extract lexical/linguistic features, rules of generation of sign sentences, transformational rules between Arabic and sign language, etc.; it can also serve for statistical translation, and this is what we are trying to do in this work by restructuring and refining the corpus (see section 3).

[1] This work was funded by a research grant from the Saudi Arabia NSTIP.

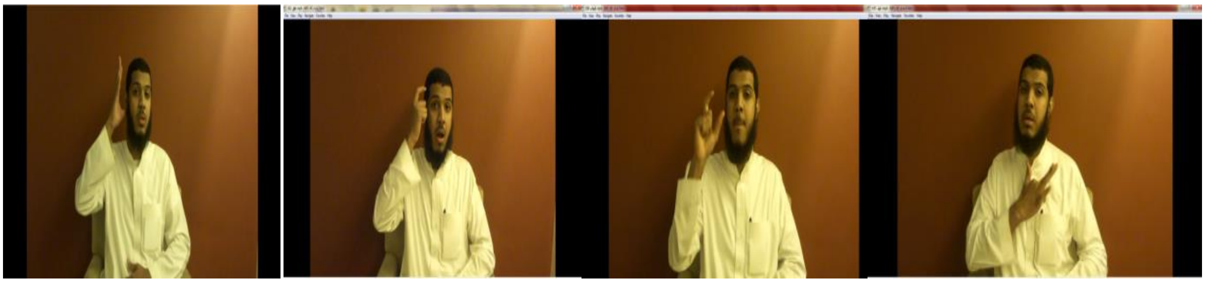


Figure 1: Extracted from video-clips of the terms “prayer”, “mind”, “Islam”, “faith” (left to write) (only a part of each one is visible in these windows).

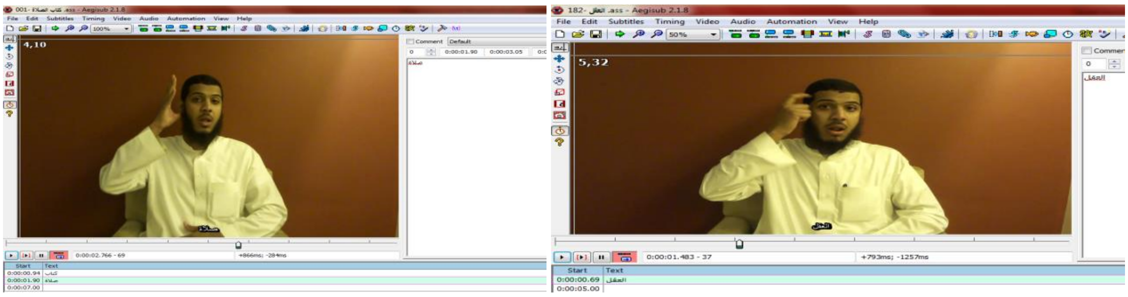


Figure 2: Alignment of the signs “prayer” and “mind” from left to right (only a part of the signs is visible in the window)

3. Refining and Restructuring the Aforementioned Corpus

In this current work, we worked out to refine and restructure the corpus to be in a suitable format for machine translation purposes. We start analyzing the corpus to extract the list of terms and their corresponding signs to construct dictionaries and their 3D animations. We have to mention here two important properties of the Sign Language to be kept in mind:

a) The Sign Language is not a direct manual representation of the oral language of the surrounding community; it has its own vocabulary and grammar that differ from the spoken language. From this fact, different words might be signed the same way, which means single sign for multiple words; this generally happens for words denoting a same concept, as for example “**أكل**” (eat) and “**طعام**” (food). For this reason, we have tried to collect for each word of our basic dictionary (previously extracted from our corpus) the other words that are signed the same way; this is done with the help (and/or validation) of the deaf group. Collecting and gathering these classes of words (we will call them “homosigns”) is extremely important for enlarging the dictionary and expanding its coverage without building new corpora and creating new 3D animations, which are hard and time consuming. This process will continue wherever a new homosign is identified, and thus continuously improving the dictionary. It is worth to mention that the coverage of the dictionary (and sign animations) is important in order to minimize the concept of finger spelling (i.e. segmenting a new word to its letters and then signing it using signs of these letters, which are already stored in the dictionary) during the translation process.

b) The Sign Language considers in general the basic forms of words (usually *stem*) and not their derivatives. From this fact, morphological variations of the same word are in most cases signed in a similar manner. Here, we have also tried to collect and group morphological variations for each word-stem of our basic dictionary (previously extracted from our corpus); this constitutes internal classes of the dictionary words.

Based on the above points, we have designed and stored our dictionary in a table with the following fields (see table 2 below):

Table 2: Structure of the dictionary table

ID	Word	Meaning (Gloss)	Stem	Morph Variations	Homo Signs	HomoSigns MorphVariations	Stem HomoSigns	Location	Sign FileName	Is Religion	Observ ations

- "**ID**": indicates the entry number of the dictionary;
- "**word**" is a word entry of the dictionary;
- "**Meaning**" is a short identification of the "word" to allow distinction between words spelling the same way but are different in meaning, which are recurrent in Arabic, such as for example "حر" (free, freeman) and "حر" (hot);
- "**Stem**" indicates the basic form of the dictionary entry "word" as the sign is related to it;
- "**MorphVariations**" represents the list of all morphological variations of the entry 'word", which are signed the same way (remember what we have explained above in point b: "word internal class");
- "**HomoSigns**" represents the list of words, which have the same sign as the entry "word" (this is what we have explained above in the point a: "word external class");
- "**HomoSignsMorphVariations**" gathers all morphological variations of the "HomoSigns" list;
- "**StemHomoSigns**" represents all stems of the "HomoSigns" list;
- "**Location**" indicates all places of occurrence of the entry "word" in the corpus to keep track of words' origins; this location is coded in this format "***topicID_partID_sentenceID_wordID***";
- "**SignFileName**" indicates the 3D animation file of the sign associated to the entry "word";
- "**IsReligion**" is a flag indicating if the entry "word" has a religious meaning or simply a normal Arabic word;
- "**Observations**" is a field to indicate particular comments relative to the entry "word".

In addition to the above table, which represents the main dictionary, we considered a complement table, in which we stored particular combination of words. The pertinence of this table comes from the fact that some sequence of words could have optimized sequence of signs, which might be different from the combination of separate word signs. So, we preferred to keep these combinations in a separate table to be used just for improvement purposes at the end of translation process; we can usually use the corresponding signs of the involved words with an acceptable level of eligibility.

Notice that we preferred to structure our dictionary in this manner to allow efficient search and lookup of sign sequences generated by the automatic translation system; this will improve the quality of the rendering component (the output of the system). This will be discussed in later stages of the work.

We have till now about 600 single entries in the main dictionary table; if we count the morphological variations and homo signs that we have already collected, we obtain around 3000 words with their 3D animations. This size will continually increase by adding new variations. The optimized combination table counts at this time around 200 entries.
We show in the figure below (figure 3) the overall ER diagram structure of the revised parallel corpus, including the sign dictionary (DictSign), topics (Topics) and their translations (Sentences), as well as the registered users and their belonging groups, and so on. These entities can be briefly described, without presenting them in details, as follows:

- **Topic:** denotes a field of the Islamic domain in the corpus, such as prayer, zakat, fasting, etc. The text corresponding to each topic is scattered in several files for practical considerations.
- **File:** represents a part of the text describing a topic of the corpus. Each file is in turn subdivided into many meaningful sentences suitable for deaf.
- **Sentence:** represents a part of a file (group of sentences) from a given topic in the corpus along them their textual and visual translations performed by the working groups.
- **Group:** indicates a set of deaf signers and a hearing interpreters, which are working with us on the preparation of the corpus (textual and visual translations). Different topics are assigned to a group for translations.
- **Comment:** is a table created for the revision of the translations; we asked other users to revise the translations performed by the working groups and to indicate their comments (if any) about the translations. These comments are used to help improving the quality of the translations.
- **Users:** is a table containing the types of users allowed in the corpus. We have four rules of a user: guest for feedback and revision, regular user for a member of the working groups, supervisor, and administrator.

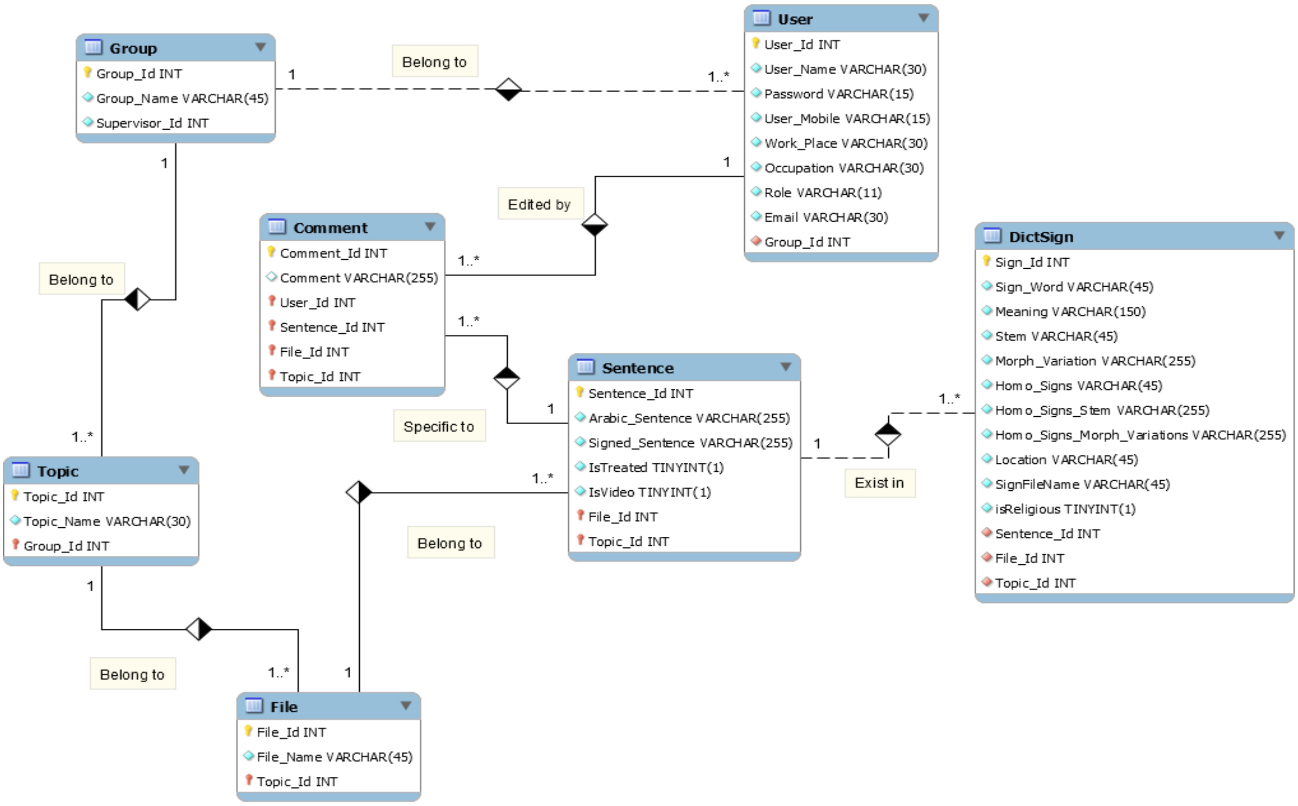
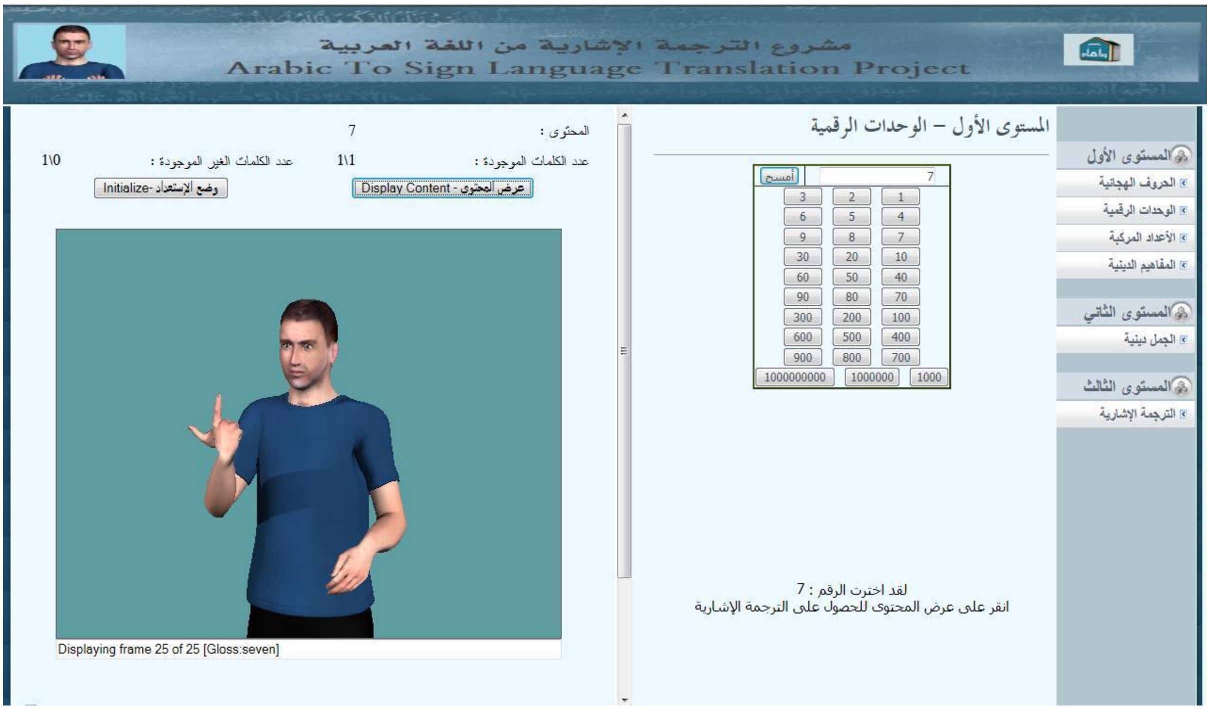


Figure 3: The overall ER diagram structure of the parallel corpus

4. Validation of the Developed Corpus

To validate the developed corpus in order to show its usability, we developed a prototype application composed of three main parts or levels to constitute a background for a future learning and teaching environment, which we aim to continuously enhance and enlarge. The first level is intended to teach some basic concepts in an incremental way (see figures 4 and 5 for some screenshots): the Arabic alphabet, the digits & numbers, and the main Islamic concepts taken from our developed religious dictionary. The second level, proposes an approach to learn Islamic subjects; it allows the user to select a field from our parallel corpus, and to interact with by choosing a content and get it translated and signed (see figure 6 for a screenshot). The third and last level, provide the user with a free-interface to enter an Arabic content and to ask for its translation; a baseline rule-based translator was developed for this purpose (see figure 7 for a screenshot). Some translation rules allowing to transit from the Arabic sentence structure to sign sentence structure were extracted from the parallel corpus and used in the rule-based translator.

Our ultimate final goal is to develop a statistical translator, which might be combined with an elaborated version of the rule-based translator to come up with an accurate translation system able to serve the deaf community for both education and communication; contents will be translated on fly and then animated and rendered using an appropriate avatar.



المستوى الأول – الوحدات الرقمية

7

المحتوى : 7

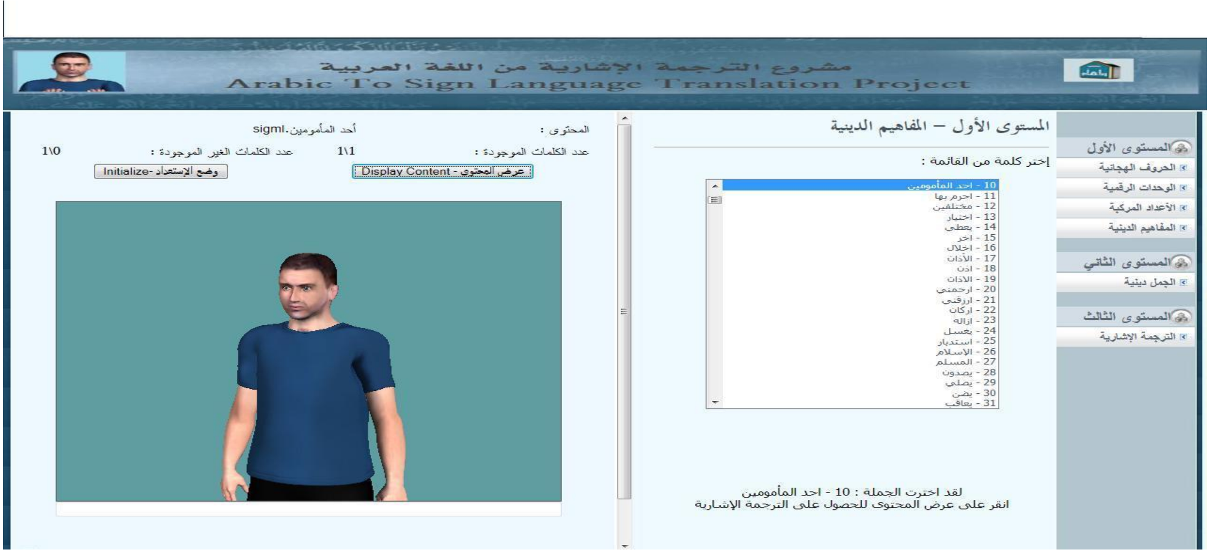
عدد الكلمات الموجودة : 1\1

عدد الكلمات الغير الموجودة : 1\0

Displaying frame 25 of 25 [Gloss:seven]

لقد اخترت الرقم : 7
انقر على عرض المحتوى للحصول على الترجمة الإشارية

Figure 4: 1st level-interface for teaching numbers


مشروع الترجمة الإشارية من اللغة العربية
Arabic To Sign Language Translation Project
المستوى الأول
الحروف الهجائية
الوحدات الرقمية
الأعداد المركبة
المفاهيم الدينية
المستوى الثاني
الجمل دينية
المستوى الثالث
الترجمة الإشارية
Display Content - عرض المحتوى
Initialize- وضع الإستعداد

المستوى الأول – المفاهيم الدينية

إختر كلمة من القائمة :

10 - احد المأمومين
11 - احرم بها
12 - مختلفين
13 - اختيار
14 - يعطى
15 - اخر
16 - اخلال
17 - الأذان
18 - اذن
19 - الاذان
20 - ارحمني
21 - ارزقني
22 - اركان
23 - ازاله
24 - يغسل
25 - استدبار
26 - الإسلام
27 - المسلم
28 - يصدون
29 - يصلي
30 - يضن
31 - يعاقب

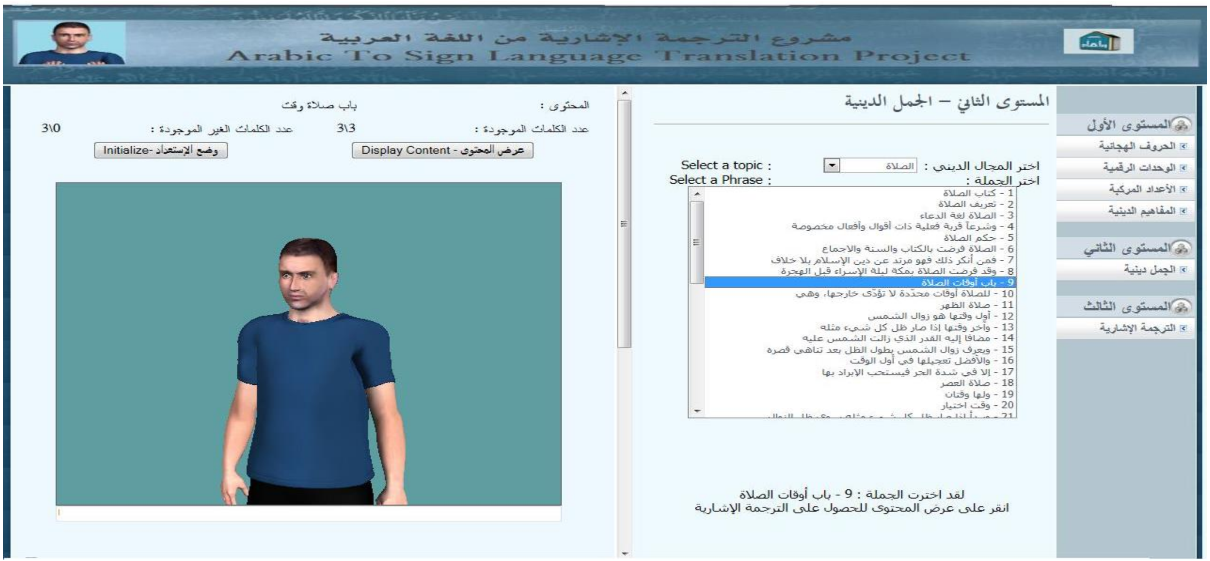
المحتوى : أحد المأمومين.sigml

عدد الكلمات الموجودة : 1\1

عدد الكلمات الغير الموجودة : 1\0

لقد اخترت الجملة : 10 - احد المأمومين
انقر على عرض المحتوى للحصول على الترجمة الإشارية

Figure 5: 1st level-interface for teaching religious concepts


مشروع الترجمة الإشارية من اللغة العربية
Arabic To Sign Language Translation Project
المستوى الأول
الحروف الهجائية
الوحدات الرقمية
الأعداد المركبة
المفاهيم الدينية
المستوى الثاني
الجمل دينية
المستوى الثالث
الترجمة الإشارية
Display Content - عرض المحتوى
Initialize- وضع الإستعداد

المستوى الثاني – الجمل الدينية

Select a topic : الصلاة اختر المجال الديني :
Select a Phrase : اختر الجملة :

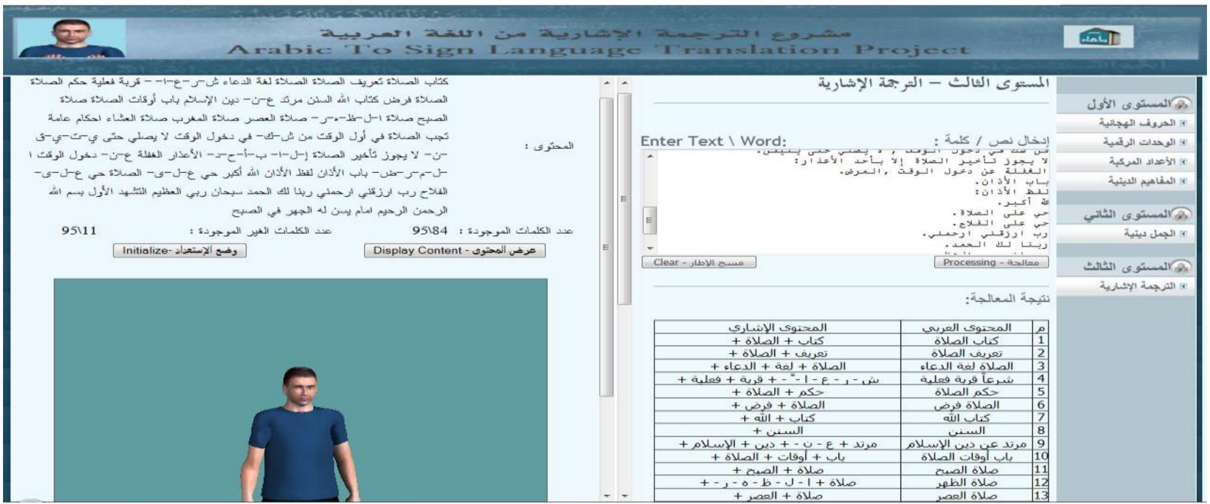
1 - كتاب الصلاة
2 - تعريف الصلاة
3 - الصلاة لغة الدعاء
4 - وشرعاً قربة فعلية ذات أقوال وأفعال مخصوصة
5 - حكم الصلاة
6 - الصلاة فرضت بالكتاب والسنة والاجماع
7 - فمن أنكر ذلك فهو مرتد عن دين الإسلام بلا خلاف
8 - وقد فرضت الصلاة بمكة ليلة الإسراء قبل الهجرة
9 - باب أوقات الصلاة
10 - للصلاة أوقات محدّدة لا تؤدّى خارجها، وهي
11 - صلاة الظهر
12 - أول وقتها هو زوال الشمس
13 - وآخر وقتها إذا صار ظل كل شيء مثله
14 - مضافا إليه القدر الذي زالت الشمس عليه
15 - ويعرف زوال الشمس بطول الظل بعد تناهي قصره
16 - والأفضل تعجيلها في أول الوقت
17 - إلا في شدة الحر فيستحب الإبراد بها
18 - صلاة العصر
19 - ولها وقتان
20 - وقت اختيار

لقد اخترت الجملة : 9 - باب أوقات الصلاة
انقر على عرض المحتوى للحصول على الترجمة الإشارية

المحتوى : باب صلاة وقت

عدد الكلمات الموجودة : 3\3 عدد الكلمات الغير الموجودة : 3\0

Figure 6: 2nd level interface for learning Islamic subjects



المحتوى : كتاب الصلاة تعريف الصلاة الصلاة لغة الدعاء ش-ر-ع-ا- - قربة فعلية حكم الصلاة الصلاة فرض كتاب الله السنن مرتد ع-ن- دين الإسلام باب أوقات الصلاة صلاة الصبح صلاة ا-ل-ظ-ه-ر- صلاة العصر صلاة المغرب صلاة العشاء احكام عامة تجب الصلاة في أول الوقت من ش-ك- في دخول الوقت لا يصلي حتى ي-ت-ي-ق-ن- لا يجوز تأخير الصلاة إ-ل-ا- ب-أ-ح-د- الأعذار الغفلة ع-ن- دخول الوقت ا-ل-م-ر-ض- باب الأذان لفظ الأذان الله أكبر حي ع-ل-ى- الصلاة حي ع-ل-ى- الفلاح رب ارزقني ارحمني ربنا لك الحمد سبحان ربي العظيم التشهد الأول بسم الله الرحمن الرحيم امام يسن له الجهر في الصبح

عدد الكلمات الموجودة : 95\84 عدد الكلمات الغير الموجودة : 95\11

المستوى الثالث – الترجمة الإشارية

Enter Text \ Word: إدخال نص / كلمة :

من شك في دخول الوقت , لا يصلي حتى يتيقن.
لا يجوز تأخير الصلاة إلا بأحد الأعذار:
الغفلة عن دخول الوقت ,المرض.
باب الأذان.
لفظ الأذان:
الله أكبر.
حي على الصلاة.
حي على الفلاح.
رب ارزقني ارحمني.
ربنا لك الحمد.

نتيجة المعالجة:

م	المحتوى العربي	المحتوى الإشاري
1	كتاب الصلاة	كتاب + الصلاة +
2	تعريف الصلاة	تعريف + الصلاة +
3	الصلاة لغة الدعاء	الصلاة + لغة + الدعاء +
4	شرعاً قربة فعلية	ش - ر - ع - ا - ً - + قربة + فعلية +
5	حكم الصلاة	حكم + الصلاة +
6	الصلاة فرض	الصلاة + فرض +
7	كتاب الله	كتاب + الله +
8	السنن	السنن +
9	مرتد عن دين الإسلام	مرتد + ع - ن - + دين + الإسلام +
10	باب أوقات الصلاة	باب + أوقات + الصلاة +
11	صلاة الصبح	صلاة + الصبح +
12	صلاة الظهر	صلاة + ا - ل - ظ - ه - ر - +
13	صلاة العصر	صلاة + العصر +

Figure 7: 3rd level-interface for translating and signing free contents

5. Conclusion

This paper is focused on the preparation of the infrastructure for building a translation system for educational purposes. An Islamic bilingual parallel corpus of Arabic and Sign languages initially developed in a first stage of our works was refined and enhanced in terms of design and structure as well as size and accessibility. With this revised corpus of around three thousand words (animated signs) dictionary and about one thousand aligned parallel translation sentences, we think that we can move to the next phase to start developing a statistical translation system. Later on, we may need to enlarge the parallel translations to ensure good performance, as the statistical approach of translation needs a large size of data to be used for training. This statistical translator will be incorporated with the translation component of the prototype we developed and presented in the paper to enhance the learning and teaching environment we looking for. The rule-based translator should also be improved and combined with the statistical one to re-enforce the translation system; this can be done by adding new translation rules and more specifications of the Sign Language sentence.

Acknowledgment

This work is funded by a research grant from the Arab Fund for Economics and Social Development.

References

Al Ameiri, F. (2011). Mobile Arabic sign language. Proc. of the International Conference for Internet Technology and Secured Transactions (ICITST), pp. 363- 367.

Elhadj, Y.O.M., Zemirli, Z.A. Ayadi, K. (2012). Development of a Bilingual Parallel Corpus of Arabic and Saudi Sign Language: Part I. Intelligent Informatics, Advances in Intelligent Systems and Computing, Volume 182, pp 285-295, Springer.

Elhadj, Y.O.M. Zemirli, Z.A., Al-faraj, B. (2012). Towards a unified 3D animated dictionary for Saudi sign language. ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 910-917, ISBN: 978-1-4503-1196-0, ACM-Digital Library.

Elhadj, Y.O.M. (2012). Multimedia Educational Content for Saudi Deaf. Neural Information Processing, Lecture Notes in Computer Science, Volume 7666, pp. 164-171, Springer.

Elhadj, Y.O.M. Zemirli, Z.A. (2014). Virtual Translator from Arabic text to Saudi Sign-Language (A2SaSL). Final technical Report, NSTIP, Al-Imam University, Saudi Arabia.

Jemni, M., Elghoul, O. (2007). An Avatar based approach for automatic interpretation of text to sign language. 9th European Conference for the advancement of the assistive Technologies in Europe, San Sebastian, 3-5 October 2007, Spain.

Jemni, M., Elghoul, O. (2008). Using ICT to teach sign language. 8th IEEE international Conference on Advanced Learning Technologies, pp 995-996, IEEE Computer Society, 2008.

Halawani, S. M. (2008). Arabic Sign Language Translation System on Mobile Devices. IJCSNS, Vol.8, No.1: 251-256, January 2008.

Mohandes, M. (2006). Automatic Translation of Arabic Text to Arabic Sign Language, AIML Journal Vol. 6, No. 4: 15-19, December 2006.