Czech-Sign Speech Corpus for Semantic based Machine Translation *

Jakub Kanis[1], Jiří Zahradil[2], Filip Jurčíček[1], and Luděk Müller[1]

[1] University of West Bohemia, Department of Cybernetics
Univerzitní 8, 306 14 Pilsen, Czech Republic
{jkanis,filip,muller}@kky.zcu.cz
[2] SpeechTech s.r.o, Pilsen, Czech Republic
jiri.zahradil@speechtech.cz

Abstract. This paper describes progress in a development of the human-human dialogue corpus for machine translation of spoken language. We have chosen a semantically annotated corpus of phone calls to a train timetable information center. The phone calls consist of inquiries regarding their train traveler plans. Corpus dialogue act tags incorporate abstract semantic meaning. We have enriched a part of the corpus with Sign Speech translation and we have proposed methods how to do automatic machine translation from Czech to Sign Speech using semantic annotation contained in the corpus.

1 Introduction

Pursuant to the law *155/1998 Sb.* the Sign Speech (SS) means Czech Sign Language and Signed Czech. The article 4 specifies the Czech Sign Language (CSE) as follows:

- Czech Sign Language is a basic communication facility of the deaf people in the Czech Republic.
- Czech Sign Language is a natural and adequate communication system. It is composed by the specific visual-spatial resources, i.e. hand shapes (manual signals), movements, facial expressions, head and upper part of the torso positions (non-manual signals). Czech Sign Language has basic language attributes, i.e. system of signs, double articulation, productiveness, peculiarity and historical dimension, and has stable lexical and grammatical structure.

The article 5 specifies the Signed Czech (SC) as follows:

- Signed Czech is an artificial language system, which facilitates communication between deaf and hearing people.

* Support for this work was provided by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416.

– Signed Czech uses grammatical resources of the Czech language, which is simultaneously loudly or unloudly articulated. The signs of the Czech Sign Language according to the individual Czech words are showed together with the articulation.

The CSE is usually used for communication between deaf people while SC is used in communication between deaf and hearing people. For example the majority of Czech TV programs for deaf people are performed in SC. Since the last time, there are programs performed in CSE too. This lack of using CSE is given by a status of the CSE before year 1989. The CSE before year 1989 was an unofficial language, using only by a deaf community. Only an oral method in combination with the SC was used for teaching deaf children. A linguistic research of sign languages started after year 1960 in the world, when W. Stockoe published his book Sign Language Structure. The Stokoe's book was the first work, which has studied the sign language from a linguistic point of view. The linguistic research of the CSE started in 90s. There exists no official form of the CSE and the language and the signs are various in different regions. And of course, like in others sign languages, there exists no written form of the CSE used by deaf people. From this reasons there is no comprehensive work about CSE syntax, but first studies show that the CSE shares some syntactic structures with others sign languages.

The using of written language instead of spoken language is wrong idea in the case of Deaf. This is because the Deaf have problems with majority language understanding when they are reading a written text. The majority language is the second language of the Deaf and its acquiring is only particular. Thus majority language translation to the sign speech is important for better Deaf orientation in the majority language speaking world. Currently human interpreters provide this translation, but their service is expensive and not always available. The machine translation systems with graphical avatar (artificial human figure) as output represent the solution, which cannot fully replace interpreters. But it can help in everyday communication.

There are two main approaches in area of machine translation (MT): linguistic and data oriented machine translation. A majority of existing translation systems is based on linguistic oriented approach, for example [1] system for translation from English to British Sign Language (BSL), [2], [3], [4] systems for translation from English to American Sign Language (ASL) and [5] system for translation from Polish to Polish Sign Language (PSL). The systems based on data oriented approach appear recently too, for example [6] statistical based system for translation from German to German Sign Language (DGS) and [7] example based system. The main problem of the data oriented approach is acquisition of training data – bilingual corpus. In this paper we describe the creation of Czech-Sign Speech corpus suitable for data oriented machine translation.

We have chosen an existing train timetable dialogue corpus (TTDC) [8] as a base of our Czech Sign Speech corpus. The choice of this corpus has a lot of advantages. Firstly, the TTDC is a record of a spontaneous telephone communication between operator and user, so the corpus covers a whole well-defined task.

Secondly, every dialog is carefully transcribed to the Czech. Thirdly, the dialogs are provided with dialog act and semantic annotation. Fourthly, the TTDC is the corpus of a telephone spontaneous speech, results acquired from its can be used in real-life and telephone applications. It opens the world of telephone communication for deaf users. Fifthly, the same corpus can be used for training of complete system translating from the spoken language to the sign language (a speech recognizer on one side and a translation system and graphical avatar on the second side). In next sections we describe the TTDC and its extension by Sign Speech translation of dialogs in detail.

2 Train Timetable Dialogue Corpus

The corpus was collected in a train timetable information center. It was recorded since April, 2000 to September, 2000. There were 6584 calls collected, from which 6353 calls (dialogues) were transcribed. Callers were mainly Czechs.

The audio part of corpus contains 106 hours of speech. Corpus uses orthographic transcription because it is more suitable for transcription of Czech spontaneous speech [9]. Spontaneous Czech contains words and usages not found either in standard written or in formal spoken Czech. From another point of view, the corpus consists of 81543 turns. Each turn starts with a speaker change. The size of the vocabulary of the whole corpus is about 12k words, and there are almost 600k tokens in it. The operator's vocabulary (5839 words) is smaller than user's vocabulary (9485 words). While a dialogue has 6 user's turns on average, the first user's turn contains 35% of user's tokens in the dialogue on average. Each turn was divided into segments that allow assigning of one dialogue act to each utterance segment.

2.1 Dialogue Act Tagging Scheme

TTDC corpus [8] is annotated by dialogue acts with additional structured semantic tags. It uses dialogue act tagging scheme slightly inspired by DAMSL (Dialogue Act Markup in Several Layer) [11] but strongly based on DATE (Dialogue Act Tagging for Evaluation) scheme [12]. The corpus uses three dimensional annotations (1) DOMAIN, (2) SPEECH-ACT, (3) SEMANTICS. The corpus has annotated whole dialogues utterances - both user's and operator's as a contrast to DATE, which was originally designed just for evaluation of dialogue systems, therefore annotation was present only at system's responses.

2.2 Data Dimensions

According to [8] TTDC tag set suppresses some disadvantages of his successors and boosts their advantages. In general, semantic annotation *normalizes* dialog utterances and therefore we believe that this annotation can help in the task of machine translation from spoken Czech to SS. We briefly describe the DOMAIN, the SPEECH-ACT, and the SEMANTIC dimensions of TTDC tag set in the following section.

DOMAIN Dimension This dimension assigns every utterance to three areas of conversational action: **Task**, **Communication**, **Frame**. The first area of DOMAIN is the task domain, which is train timetable inquiry answering. The second area is managing communication channel, it manages the verbal channel and provides evidence what has been understood. Finally, the third area is a situation frame, which refers to an apology or an instruction contained in a sentence. This domain is not as frequent in human-human dialogs as in human-machine dialogs.

Automatic sign language translation can use domain information to focus on task sentences and handle the communication problems in correct sign-specific form.

SPEECH-ACT Dimension This dimension refers to an utterance's communicative goal, independently on an utterance form. This dimension differentiates utterances that have the same value of the SEMANTICS dimension. For instance, the SPEECH-ACT dimension values REQUEST-INFO and PRESENT-INFO can refer to the same value in the SEMANTICS dimension, e.g. DEPARTURE(TIME, FROM(STATION)). TTDC scheme use namely these speech acts: request-info, present-info, verify, verify-neg, offer, acknowledgment, status-report, explicit-confirmation, implicit-confirmation, instruction, apology, opening, closing and speech-repair.

In SS translation domain, we are planning to use extracted information about utterance segments to for example sentence type resolution.

SEMANTIC Dimension This dimension captures task relevant information from each utterance. In train timetable inquiry answering task domain; the goal of communication is to determine information needed to answer an inquiry, e.g. a departure train station or time of desired departure. In the sentence "Is there any train to Pilsen at eight am", the semantics is REQUEST=departure, TO=Pilsen, and TIME=eight am. The extracted semantic concepts should be also sufficient for machine translation. Semantic annotation has preserved the hierarchical structure of an utterance, but it stills prevailed simplicity. Although, the semantic layer generalizes sentences, this generalization is precious and because of vocabulary reduction it makes machine translation process simpler. Another possibility is to conditioning translation with respect to semantic annotation.

There are two main semantic concepts defined in TTDC. DEPARTURE is a concept for an utterance that represents question about departure of a particular train (answer is usually exact time when the train leaves a particular train station) and ARRIVAL is similar concept for arrival to particular station. Each of previous semantic concepts is allowed to have 27 non-terminal leaves (concepts): FROM, TO, THROUGH, IN_DIRECTION, TRAIN_TYPE, TIME, and few rare concepts: BACK, DELAY, DISTANCE, DURATION, GREETING, PRICE, PERSON, AREA, WAIT, etc. Nearly all concepts can be nodes in hierarchical semantic tree, as there are very weak constraints on their possible relations in natural spoken language.

Totally, we have 1000 dialogues semantically (manually) annotated, that means 16645 dialogue acts and 1202 of them are unique. The corpus consists of totally 26472 semantic tokens (concepts) with hierarchical binding. See annotation sample in Table 1.

Table 1. A sample: part of dialogue including SC translation and semantic annotation.

Speaker	DA Semantics	Czech Sentence SC Translation
operator	comm,opening NIL	informace prosím *informace/1 _/2*
user	comm,opening NIL	dobrý den *dobrý_den/1,2*
	task,request-info DEPARTURE(TIME, TRAIN_TYPE, TO(STATION))	já mám prosbu jakpak jedou dneska osobní vlaky ňák dopoledne do starýho_plzence *já/3 potřebovat/4,5 kdy/6 jet/7 dnes/8 osobní_vlak/9,10 _/11 dopoledne/12 do/13 starý/14 plzeň/14 malý_věc/14*
operator	frame,status-report NIL	no tak tam už moc na výběr nemáte *_/1 _/2 _/3 už/4 moc_hodně/5 _/6 výběr/7 ne/8*
	task,present-info TIME, TIME	tedka jede v osm šestnáct jestli stihnete potom až v jedenáct deset *teď/9 jet/10 v_ve/11 osm_hodin/12 šestnáct/13 jestli/14 stihnout/15 potom/16 až/17 v_ve/18 jedenáct_hodin/19 deset/20*
user	comm,implicit-conf TIME	až v jedenáct deset *až/1 v_ve/2 jedenáct_hodin/3 deset/4*
	task,acknowledgment ACCEPT(TIME, FROM(STATION))	a to by tak ňák stačilo těch jedenáct deset z hlavního *_/5 _/6 _/7 _/8 _/9 stačit/10 _/11 jedenáct_hodin/12 deset/13 z_ze/14 důležitý/15*
	task,request-info VERIFY(TRAIN_TYPE)	jo a dá se tam vzít kočárek *_/16 _/17 moci/18 _/19 tam/20 vzít/21 _/22*

3 Sign Speech Translation

3.1 Process of Translation

There exist no formal written forms of the SS, thus the main problem in SS corpus building is a choice of the appropriate written forms. In the first stage of corpus building we decide to extend the TTDC corpus by Signed Czech translation of dialogues. The SC sentence has the same grammatical structure like Czech sentence and uses the signs of CSE corresponded to the individual Czech words. The SC sentence in written form can be simply represented by a sequence of CSE signs. Every CSE sign is represented by a unique string. To speed up the manual translation process we have extended the annotation tool DAE, proposed in [8]. This software, including our extension, is distributed under GPL license and is available for download at project webpage [13].

Every translator uses the same CSE dictionary to ensure a consistence of translations. We use a text version of the most extensive CSE dictionary [14] (this dictionary contains 3063 signs). We have added two special signs into the dictionary. The first is used in the case that some Czech word is not translated

in the corresponded SC sentence. And the second is used for the words, which need to be finger-spelled. This dictionary is a part of our annotation tool DAE. And the translator can choose only the signs from this dictionary in translation process of dialogues.

We use an explicit alignment too. The translator has to match every Czech word with one or more signs in SC sentence. The one sign can be match with more Czech words too. For example SC sentence (English literal translations of original SC sentence): *"good_morning I need when go regional_train to old pilsen small_thing"* corresponds to Czech sentence: *"good morning I have a question how can I go today by regional train to old_pilsen"* . Here on one hand the Czech words *good* and *morning* correspond to one sign *good_mornig* and other hand the word *old_pilsen* corresponds to three signs *old* , *pilsen* and *small_thing* . The explicit alignment has some advantages. We can simply check if the translator translates all words from Czech sentence (i.e. every Czech word has to be assigned at least one sign). And we can straightly create a bilingual dictionary, which is phrase based (one or more Czech words can corresponds to one or more signs).

3.2 Direct Translation System based on Explicit Alignment

The bilingual dictionary with phrases can be used as a simple direct translation system. If there are more possible translations for one word/phrase we choose the most probable possibility. We have collected 800 dialogues in SC since May 2006. We have used 720 dialogues for dictionary creation and the rest 80 dialogues for testing. The statistical data and results are in Table 2.

Table 2. The result of direct translation system

	Training data	Testing data
no. of sentences	10241	1188
no. of distinct words	3557	1019
no. of distinct signs	658	368
no. of running words	-	8122
no. of OOV words	-	221(2.72%)
SER[%]	-	50.5
WER[%]	-	14.0

Where SER is sentence error rate, it is a ratio of bad sentence translations to a number of all translated sentences. And WER, word error rate, is similarly a ratio of bad word translations to a number of all translated words.

3.3 Semantic based Machine Translation

The dialog act and semantic annotation of TTDC corpus can be used in different ways for machine translation. Firstly, this annotation can be considered to be an

Interlingua for Czech and Sign Speech. The interlingual representation of text is independent of a source language. How we can see in Figure 1, there is the same semantic tree for Czech sentence and its SC translation. The MT system then works as follows: the source language text is converted to the interlingual representation first and then the target language text is generated from this language-independent, interlingual representation. Secondly, the SPEECH-ACT dimension of dialog act annotation can be used for the sentence type resolution. For example in CSE is important to distinguish if the question is yes/no- or wh-type of question, because every type uses other non-manual signals.

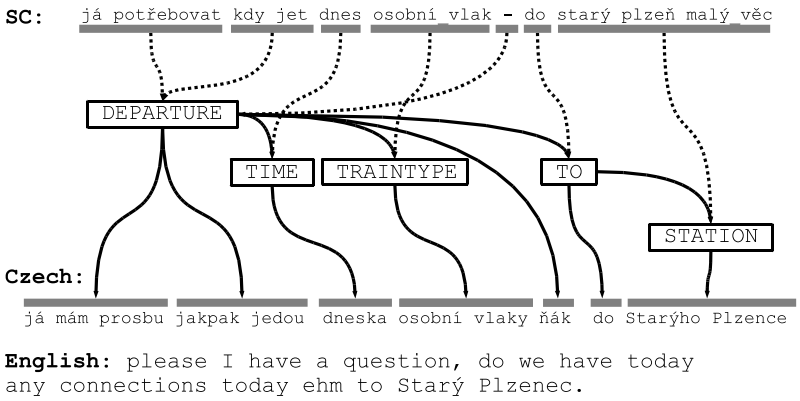

Fig. 1. Semantic annotation of Czech and SC sentence

4 Conclusions and Future Work

In this paper we have described the first stage of a Sign Speech corpus building and a simple direct translation system based on phrase bilingual dictionary. The SS corpus is based on the existing TTDC corpus. The TTDC corpus is a dialogue corpus with a dialog act and a semantic annotation. In the first stage of SS corpus building we have added the Signed Czech translation of dialogues. To speed up the manual translation process we have extended the annotation tool DAE, proposed in [8]. Every translator uses the same CSE dictionary [14] with two special signs added (signs for 'no translation' and 'finger-spelling'). Every translator has also to decide the explicit alignment between a Czech sentence and the SC translation. We have created a simple translation system based on this explicit alignment. The sentence error rate of proposed system is 50.5 %. This quite good result is given mainly by a strong linguistic similarity of both languages (SC uses the grammatical resources of Czech).

We plan to add CSE dialogues translations to the corpus in the second stage of the corpus building. We can use the same CSE dictionary and annotation tool (with necessary modifications). The written form of CSE will be more complicated than the written form of SC. Especially if we want to describe a spatial component of CSE. Our main goal is to design the CSE written form, which would be suitable for CSE synthesizer.

References

1. Marsahll, I., Safar, E., "Sign Language Generation using HPSG", In Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-2002, Japan. 2002.
2. Speers, dA.L., "Representation of American Sign Language for Machine Translation", PhD Dissertation. Department of Linguistics, Georgetown University.
3. Zhao, L. et al., "A Machine Translation System from English to American Sign Language", Association for Machine Translation in the Americas. 2000.
4. Huenerfauth, M., "A Multi-Path Architecture for Machine Translation of English Text into American Sign Language Animation", In Proceedings of the Student Workshop at the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2004). Boston, MA, USA. 2004.
5. Suszczanska, N., Szmal, P., Francik, J., "Translating Polish Texts into Sign Language in the TGT System", 20th IASTED International Multi-Conference Applied Informatics AI 2002. Innsbruck, Austria. 2002, s. 282-287.
6. Bungeroth, J., Ney, H., "Statistical sign language translation", In: Streiter, Oliver / Vettori, Chiara (eds): LREC 2004, Workshop proceedings : Representation and processing of sign languages. Paris : ELRA (2004) - pp. 105-108
7. Morrissey, S., Way, A., "An Example-Based Approach to Translating Sign Language", 2nd International Workshop on Example-Based Machine Translation - At MT Summit X. 2005
8. Jurčíček, F., Zahradil, J., Jelínek, L., "A human-human train timetable dialogue corpus", In Interspeech Lisboa 2005. Bonn : ISCA, 2005. s. 1525-1528. ISSN 1018-4074.
9. Psutka, J., Ircing P., Hajic, J., Radova, V., Psutka, J.V., Byrne, W., and Gustman, S., "Issues in annotation of the Czech spontaneous speech corpus in the MALACH project", Proceedings of the International Conference on Language Resources and Evaluation, LREC, 2004.
10. Young, S., "Talking to Machines (Statistically Speaking)", Proceedings of the International Conference on Spoken Language Processing, Denver, USA, 2002.
11. Allen, J. and Core, M., "DAMSL: Dialog Act Markup in Several Layer", http://www.cs.rochester.edu/research /cisd/resources/damsl, 1997.
12. Walker, M., and Passonneau, R., "DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems", IEEE Trans. Speech and Audio Proc., 7(6):697–708, 1999.
13. Jurčíček F., Kanis J., Zahradil, J., "DAE, Annotation tool project page", http://ui.zcu.cz/projects/dae/, 2006
14. Langer, J., Ptáček, V., Dvořák, K., "Znaková zásoba českého znakového jazyka", Univerzita Palackého v Olomouci, Olomouc 2004.