The KIT Lecture Corpus for Speech Translation

Sebastian Stüker[1,2], Florian Kraft[1], Christian Mohr[1], Teresa Herrmann[1], Eunah Cho[1] and Alex Waibel[1]

[1]Interactive Systems Laboratories
[2]Research Group 3-01 'Multilingual Speech Recognition'
Karlsruhe Institute of Technology
Karlsruhe, Germany
`{sebastian.stueker|florian.kraft|christian.mohr|teresa.herrmann|eunah.cho|`
`alexander.waibel}@kit.edu`

Abstract

Academic lectures offer valuable content, but often do not reach their full potential audience due to the language barrier. Human translations of lectures are too expensive to be widely used. Speech translation technology can be an affordable alternative in this case. State-of-the-art spoken language translation systems utilize statistical models that need to be trained on large amounts of in-domain data. In order to support the KIT lecture translation project in its effort to introduce speech translation technology in KIT's lecture halls, we have collected a corpus of German lectures at KIT. In this paper we describe how we recorded the lectures and how we annotated them. We further give detailed statistics on the types of lectures in the corpus and its size. We collected the corpus with the purpose in mind that it should not just be suited for training a spoken language translation system the traditional way, but should also allow us to research techniques that enable the translation system to automatically and autonomously adapt itself to the varying topics and speakers of the different lectures.

Keywords: speech translation, talk translation, corpus

1. Introduction

Academic lectures and technical talks often provide high quality content that is of value to audiences that have many different mother tongues. But many lectures often do not reach their full potential audience due to the limits imposed by the language barrier between lecturer and potentially interested listeners.

(**?**)

While, in principal, simultaneous translations by human interpreters are a solution to bridge the language barrier, in reality this approach is too expensive for most of the many lectures held every day across the world. Besides the problem of high costs, human interpreters would also not be available in sufficient numbers to service all needs.

Here, technology in the form of *spoken language translation* (SLT) systems can provide a solution, making lectures available in many languages at affordable costs. Therefore, one of our current research focuses is the automatic translation of speeches and academic lectures (Fügen et al., 2007)(Fügen, 2008).

Within our research we are interested in two principal scenarios. The first scenario is the *simultaneous* translation of lectures as they happen. For this, it does not matter to us whether the audience is present in the lecture hall or possibly remotely connected by modern means of, e.g., livestreaming or telepresence systems.

The second scenario is the *offline* translation of recorded lectures, e.g., stored in databases to be viewed by individuals at later times. Such databases of, as of now, untranslated lectures are increasingly being created by universities world wide. As educational institutions start to offer their lectures, or at least parts of them, on-line, these databases can be increasingly found on the *World Wide Web*

(WWW). For example, the Massachusetts Institute of Technology (MIT) offers all courses online through their *OpenCourseWare* (OCW) web site[1] (Atkins et al., 2007), while Carnegie Mellon University makes its lectures available through *Carnegie Mellon Online lectures* (Thille, 2008). Apple has created the *iTunes U* service[2] in order to distribute educational audio, video, and PDF files. As of now over 75,000 files from universities and institutions from over a dozen countries are available on it.

In addition, conference talks and other lectures, that are very close in style and content to traditional university lectures, are also increasingly available from the WWW. For example, TED makes its lectures and talks available via their website[3]. Another example is http://videolectures.net/ which calls itself a free and open access educational video lectures repository and as of today hosts over 90,000 lectures.

1.1. Lecture Translation at KIT

At the *Karlsruhe Institute of Technology* (KIT) most lectures are held in German. This is often a significant obstacle for students from abroad wishing to study at KIT, as they need to learn German first. In order to be able to truly follow the often complex academic lectures, the level of proficiency in German that the foreign students need to reach is quite high.

We therefore pursue the goal of aiding those students, by bringing simultaneous speech translation technology into KIT's lecture halls.

Current state-of-the-art spoken language translation sys-

---

[1]http://ocw.mit.edu/index.htm
[2]http://www.apple.com/de/education/itunes-u/
[3]http://www.ted.com/

tems make use of statistical models that need to be trained on large amounts of in-domain training data (Federico et al., 2011) (Lamel et al., 2011) (Boudahmane et al., 2011). Therefore, in order to tailor our simultaneous speech translation system to KIT wide lectures, we started a comprehensive data collection effort for lectures at KIT.

2. Data Needs of Spoken Language Translation Systems for Lectures

Academic lectures and conference talks have only recently become the interest of research projects and evaluations, such as the IWSLT evaluation campaign (Paul et al., 2010) (Federico et al., 2011) or the lecture translation project at KIT. As discussed above, here, SLT systems can service a need that cannot be fulfilled by human translators. While the quality of SLT systems still lacks that of human translators, their performance has increased over the years, and they start to be of value to users now (Hamon et al., 2007). In order to develop SLT systems that perform well for the task of translating German KIT lectures, new data resources are needed that are not available so far.

While the concentration on lectures already limits the domain of the SLT system somewhat, the topics encountered in the lectures can still be arbitrary and thus, the domain of lectures can still be seen as practically unlimited. Data resources made available for the speech translation systems therefore need to capture this diversity in order to be able to train appropriate models, and, even more important, to reliably assess the quality of the systems developed.

As for all learning systems that use statistical models, speech translation systems need two separate sets of resources: one for training and one for evaluating the system and monitoring its progress.

For training and testing the *automatic speech recognition* (ASR) component of the SLT system large amounts of in-domain audio data are needed that are transcribed at word level. For the *machine translation* (MT) component of the system, data is needed that consists of parallel sentences in the required domain in all languages between which the system is supposed to translate.

Since lectures provide such a diverse set of topics, we anticipate that—especially with respect to the language model, translation model, and vocabulary—the traditional approach of training systems on a fixed set of data and then deploying them, will not be sufficient. We assume that reasonable performance will only be reached by systems that are able to flexibly and autonomously adapt themselves to varying topics of lectures. In order to facilitate this adaptation process, the presence of verbose meta-data, such as the name of the lecturer, his field of expertise, the title of the lecture, or the slides used by him, is very valuable.

The corpus collected by us reflects those needs, and is thus not only intended as a resource for training SLT systems the traditional way, but also as a tool for conducting research to advance the state-of-the-art in autonomous and unsupervised adaptation for SLT systems.

3. Corpus Collection

We started the collection of our German lecture corpus with two main goals in mind. First, we wanted to obtain the nec-essary resources to train a speech translation system on the German lecture domain the traditional way, by annotating sufficient amounts of in-domain data. But since this traditional approach is not suited for deploying lecture translation systems for all lectures at KIT, we wanted to collect a database that could be used for developing the necessary technology to be able to let the translation system adapt itself automatically to new lecturers and topics.

We therefore sought to collect a wide variety of lectures from as many different faculties, lecturers, and topics as possible. By collecting not only single lectures, but where possible, a series of lectures from the same class, we sought to get the necessary base to perform experiments that simulate the development of the lecture translation system over time when applied in real-life.

3.1. Data Collection and Annotation

Our data collection was performed within the lecture halls at KIT. The diversity of the collected lectures evolved over time. While in the beginning we collected lectures given by people from our laboratory, we later expanded the collection to lectures from the computer science faculty in general, and then, later on, to lectures from all faculties of KIT. In addition to just collecting the audio from the speaker, we also collected, where possible, the slides used in the lecture, including timing information of slide changes, and made a video recording of the lecture. The lecturer's audio was then manually transcribed and translated in order to create the necessary corpus.

3.1.1. Recording

The recordings of the lectures were conducted by student part timers that were trained on the recording equipment by staff from our laboratory. Figure 1 gives an overview of the recording process.

The lecturer's speech was picked-up with three different microphones: a) a close-talking microphone (Countryman E6 Earset), b) a lavalier microphone (Sennheiser ME4N), and c) a dynamic hand held microphone that is worn around the neck with a string. The last type of microphone is the one normally used in KIT's lecture halls to broad cast the lecturer's speech over the lecture hall's PA system. Depending on the lecture hall and lecture recorded the last type of microphone was not available in all cases. All microphones are wireless microphones as not to inhibit the movement of the lecturers. While the close-talking microphone and the lavalier microphone are connected to a wireless transmitter, the dynamic hand held microphone has an inbuilt transmitter. The recording student set up the transmission frequency of wireless senders and receivers (Sennheiser EW 100 G3) so as not to interfere with other wireless transmissions in the environment. The receivers of the three wireless microphones were connected to multi channel sound cards (Cakewalk UA-25EX). One channel (usually from the lavalier microphone) was forwarded to the audio-input of the camcorder (Canon Legria HFS20E), in order not to capture the environmental noise of the lecture hall in the video recording. The multi channel sound cards were connected via USB to a recording laptop, where each channel was stored at 48 kHz with 24 bit resolution.
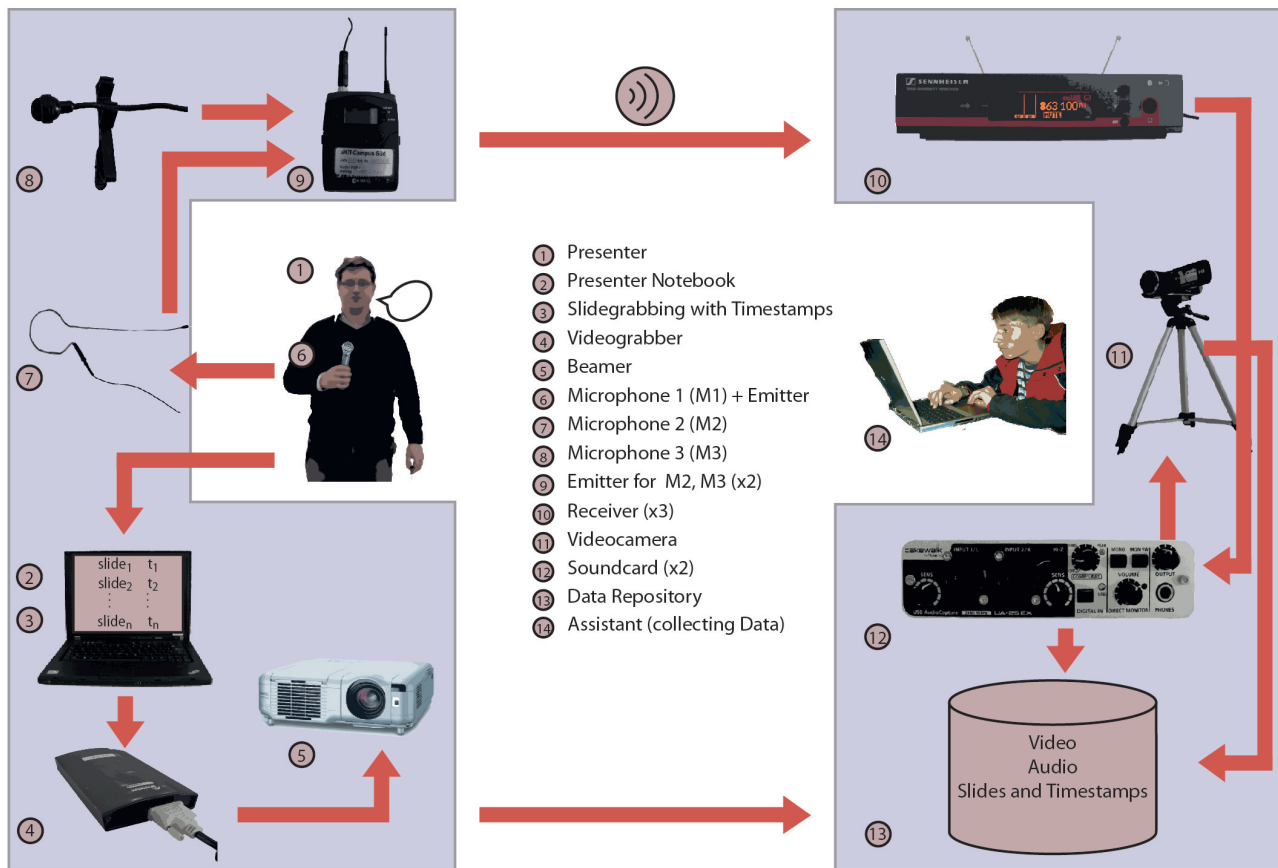
Figure 1: Overview of the lecture recording process, using three wireless microphones, recording of the slides' video with timestamps, as well as video recording of the presentation

The recording student adjusted and monitored the gain level and audio quality of all devices and channels in the transmission chain using headphones and visual feedback during the whole lecture.

In addition to the audio and video from the lecturer we also collected his slides. In the beginning we simply copied the slides from the lecture. Later on we used a tool on the presenters laptop that captures the slides and the transition times for PowerPoint slides. When it was not possible to install the tool on the lecturer's laptop, the recording assistant logged the slide changes manually using the same software on his laptop. Since the presentation format of different lecturers or even in one and the same lecture might vary, and we wanted to become independent of the specific presentation software used, we started capturing the video signal send from the presentation laptop to the video projector using an external video grabber (Epiphan DVI2USB). Slide transition times as well as text content from the slides can now be extracted offline in an automatic way using OCR.

After each lecture the audio recordings were post-processed by the recording assistants. Their amplitude was normalized and they were down sampled to 16 kHz with 16 bit resolution, which is the standard input format for our speech recognition front-end.

All captured and processed information including the various acoustic and video channels, the time stamps and meta-information about the lecturer, the lecture time and place and the recording protocol are stored and documented in a predefined structure.

3.1.2. Transcription

As soon as newly processed lecture recordings were available, the recording channel with best audio quality was added to the transcription chain. The real-time factor for creating a one-pass transcription is between 15x and 30x depending on the experience and carefulness of the transcriber. There is actually not too much variability in the acoustic events across different channels of the same recording, so the transcription can be applied to all microphone channels which are time synchronous. The transcriptions of the lecture recordings were conducted by student part timers that were trained by a linguist or by an experienced transcription student. The transcription guidelines were derived from the transliteration guideline for spontaneous speech data developed within Verbmobil II (Burger, 1997). The following changes were applied:

- German umlauts were directly written down.

- Comments to pronunciation, e.g., slang, were marked with a percent sign appended to the original form (haben%) instead of using arrow brackets

- Spellings and acronyms were written without the dollar signs as marker.

- The tilde sign was omitted as a marker for named entities.

- The hash sign was omitted as a marker for numbers.

- Numbers were written without a slash and separated in 10th groups (i.e. "vierundfünfzig", but "ein hundert vierundfünfzig")

- Overlapping speech, background speech, and background noises are not annotated.

The transcription is performed in three stages. In the first stage a transcriber segments and transcribes the recording. In the second pass a different transcriber improves the transcripts of the first pass. In the third pass the output of the second pass is spell and sanity checked to account for new German spelling rules as well as typing errors.

3.1.3. Translation

After transcription the German lectures are translated into English and French. The translations are performed by student research assistants who study at the School of Translation, Interpreting, Linguistics and Cultural Studies of the University of Mainz in Germersheim, Germany. As translators in training they produce high-quality translations for the lecture transcriptions which present a challenging task containing on the one hand very specialized technical terms and on the other hand spontaneous and colloquial speech. We make sure that either the source or target language is the translator's mother tongue. All translators work part-time and on an hourly basis as their study programme allows.

4. Corpus Details and Transcriptions

Our collection efforts already started in 2006, but have picked up pace over the last two years. While recording and annotation is still on-going we will give in the following paragraphs an overview of the current status of the size and properties of the corpus as of spring 2012.

Table 1 gives an overview of the types and amounts of lectures recorded. One can see that the bulk of the lectures comes from computer science. This is due to the fact that in the beginning we exclusively collected computer science lectures and only two years ago started to also collect lectures from other departments. In addition to real university lectures, we also occasionally collected speeches and addresses given by university representatives, e.g., the president of the university or the dean of a specific faculty, at festive or public events.

As transcription is a time-intensive process, one can see that the majority of the recorded data is not transcribed yet. But at least within computer science, the amount of ca. 46h of completely transcribed and 29h of at least first pass transcribed lectures is already significant and suited for traditional system development.

The amount of translated material is naturally lacking behind the amount of transcribed material, as it has to wait for the transcriptions first. The amount of material translated to French is particularly small, as we were only able to recruit one translator for French so far.

Table 2 gives some more statistics on the lectures. The recorded lectures contain a total of 44 speakers, the majority of them, 36, being male speakers. Most lectures, 38, are from the computer science department, though these lectures also contain lectures from the so-called Center of Applied Law which teaches law with a focus on aspects related to information technology, and is thus officially part of the computer science department. From non-computer science departments mostly only one lecture each has been recorded so far.

In case we were able to record at least five lectures of a class, we call this a series of lectures, otherwise we talk of single lectures. Most of the recordings of the lectures are single lectures instead of a series of lectures from the same class. This is often due to the fact, that many lecturers agreed to have one or two lectures recorded, but thought the recording process to be too intrusive as to have the complete class recorded.

Table 1: Duration of the recorded lectures and word count statistics on the transcriptions and translations

Type of talk recorded	(hh:mm:ss)	# words
Computer Science (CS)		
Total	185:21:09	-
2nd pass transcribed	45:47:09	438,922
1st pass transcribed	29:03:50	273,281
not transcribed	110:30:10	-
English translation	29:39:26	290,610
French translation	2:00:36	20,878
Non Computer Science (Non-CS)		
Total	7:52:16	88,214
2nd pass transcribed	5:35:53	55,575
1st pass transcribed	2:16:23	32,639
English translation	3:47:25	43,301
French translation	0:05:47	777
Miscellaneous Talks (Misc)		
Total	4:13:13	-
2nd pass transcribed	3:38:06	30,328
not transcribed	0:35:07	-
English translation	3:38:06	30,328

5. Initial Experiments

In order to give an impression of how difficult the collected lectures are for the task of spoken language translation, we performed some initial recognition and translation experiments on selected lectures from our corpus. For these experiments we used a legacy speech recognition and a legacy translation system and concatenated them.

5.1. Test Data Selection

From the available lectures we selected two of the miscellaneous talks and three of the computer science lectures. One of the CS lectures and one of the Misc talks were given by the same speaker. The lectures were chosen, because

Table 2: Statistics on speakers and lecture types

General	
Number of speakers	44
Female speakers	8
Male speakers	36
Departments	
Computer Science	38
Mechanical Engineering	2
Electrical Engineering	1
Civil Engineering, Geological and Environmental Science	1
Humanities and Social Science	1
Lecture series recorded	9
Single lectures recorded	31
Miscellaneous talks recorded	9

Table 3: Initial Experiments—ASR Results: word error rates (WER) in %, language model perplexity (LM PPL), and out-of-vocabulary rates (OOV) in %

Lecture	WER	LM PPL	OOV
Misc 1 spk 1	20.7	289.4	1.2
Misc 2 spk 2	26.6	214.2	0.8
CS spk 1	29.8	311.6	1.0
CS spk 3	43.4	305.7	1.4
CS spk 4	36.4	291.1	1.3

transcriptions and translations were available for them, and because they were not included in the training data of the machine translation system used, which has already been trained, at least in part, on the newly collected data.

5.2. ASR Results

For performing our ASR experiments we used our German speech-to-text system with which we participated in the 2011 Quaero evaluation (Lamel et al., 2011). This system is a further development of our 2010 evaluation system (Stüker et al., 2012). The domain targeted within the Quaero evaluation is that of broadcast news, broadcast conversation and podcasts downloaded from the Web. Therefore, there is a substantial mismatch in domain to which the Quaero ASR system has been tuned and the targeted lecture domain.

Table 3 shows the case-sensitive word error rates (WER) for the five recordings in the test set. It further shows the perplexity (PPL) of the ASR system's language model on the lectures and talks, as well as the out-of-vocabulary (OOV) rates of the system's vocabulary.

One can see that recognizing the miscellaneous talks is significantly easier with the Quaero system than recognizing the computer science lectures. This is both reflected in the lower word error rates and language model perplexity. The OOV rate on all talks is considerably low, which at least in part is due to the large vocabulary (300k) which makes use of sub-word units.

5.3. Speech and Machine Translation Results

Two types of machine translation experiments were conducted on the lecture data. For one, the output of the automatic speech recognition system as described above was used as input to an SMT system. Before translation text normalization, compound splitting and smart-casing was applied to the ASR output. A second set of experiments, where we translated the reference transcriptions of the lectures shows the upper bound of translation quality. For translation we used a state-of-the-art phrase-based SMT system especially developed for the translation of lecture data. Translation and language models were adapted using a large amount of in-domain lecture data, which we gathered from lectures given at KIT. We also utilized TED[4] data as an additional in-domain training data, for a better adaptation towards speech translation. To cover long-range reorderings between the source and target language, part-of-speech-based reordering was applied. Table 4 presents the results of the end-to-end evaluation (ASR) and the oracle evaluation (Transcript) of this first automatic speech translation experiment. The translation quality was measured ignoring punctuation and is presented using the case-insensitive BLEU score (Papineni et al., 2002).

Table 4: Initial Experiments—MT Results: BLEU score on ASR input and reference transcriptions

Lecture	ASR	Transcripts
Misc 1 spk 1	25.30	33.53
Misc 2 spk 2	24.36	34.90
CS spk 1	21.02	30.19
CS spk 3	13.35	26.57
CS spk 4	18.48	33.01

6. Conclusion

In this paper we presented our efforts in collecting a corpus for supporting our research within the frame work of the KIT lecture translation project. The goal of the project is to bring spoken language translation technology into KIT's lecture hall, in order to make it easier for foreign students, that are not sufficiently fluent in German, to attend lectures at KIT. While originally seeded in the computer science department, our corpus is intended to be diverse enough, in order to reflect the wide diversity of topics and speaker encountered in lectures at KIT. The corpus is not only intended to enable us to train spoken language translation systems in the traditional way, but rather to support research in advancing techniques that will allow the translation system to automatically and autonomously adapt itself to the varying topics and speakers.

7. Acknowledgements

---

[4] http://www.ted.com

8. References

Daniel E. Atkins, John Seely Brown, and Allen L. Hammond. 2007. A review of the open educational resources (oer) movement: Achievements, challenges, and new opportunities. Technical report, The William and Flora Hewlett Foundation, February.

Karim Boudahmane, Bianka Buschbeck, Eunah Cho, Joseph Maria Crego, Markus Freitag, Thomas Lavergne, Herrmann Ney, Jan Niehues, Stephan Peitz, Jean Senellart, Artem Sokolov, Alex Waibel, Tonio Wandmacher, Jörn Wübker, and François Yvon. 2011. Speech recognition for machine translation in quaero. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA.

Susanne Burger. 1997. Transliteration spontanprachlicher Daten - Lexikon der Transliterationskonventionen - Verbmobil 2. http://www.phonetik.uni-muenchen.de/Forschung/Verbmobil/VMtrlex2d.html.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine Translation*, 21:209–252.

Christian Fügen. 2008. *A System for Simultaneous Translation of Lectures and Speeches*. Ph.D. thesis, Universiät Karlsruhe (TH), November.

O. Hamon, D. Mostefa, and K. Choukri. 2007. End-to-end evaluation of a speech-to-speech translation system in TC-STAR. In *Proc. of Machine Translation Summit XI*, pages 223–230, Copenhagen, Denmark, 10-14. September.

Lori Lamel, Sandrine Courcinous, Julien Despres, Jean-Luc Gauvain, Yvan Josse, Kevin Kilgour, Florian Kraft, Le Viet Bac, Hermann Ney, Markus Nubaum-Thom, Ilya Oparin, Tim Schlippe, Ralf Schlter, Tanja Schultz, Thiago Fraga Da Silva, Sebastian Stüker, Martin Sundermeyer, Bianca Vieru, Ngoc Thang Vu, Alexander Waibel, and Ccile Woehrling. 2011. Speech recognition for machine translation in quaero. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2–3. December.

Sebastian Stüker, Kevin Kilgour, and Florian Kraft. 2012. Quaero speech-to-text evaluation systems. In *High Performance Computing in Science and Engineering '11*, pages 607–618. Springer Berlin Heidelberg.

Candace Thille. 2008. Building open learning as a community-based research activity. In Toru Iiyoshi and M.S. Vijay Kumar, editors, *Opening Up Education— The Collective Advancement of Education through Open Technology, Open Content, and Open Knowledge*. The Carnegie Foundation for the Advancement of Teaching, The MIT Press, Cambridge Massachusetts, USA and London, England.