*Online Journal of Applied Knowledge Management*
A Publication of the International Institute for Applied Knowledge Management

*Volume 2, Issue 1, 2014*

# Can automatic speech recognition be satisficing for audio/video search? Keyword-focused analysis of Hebrew automatic and manual transcription

**Vered Silber-Varod,** The Open University, vereds@openu.ac.il

**Nitza Geri,** The Open University, nitzage@openu.ac.il

## Abstract

*With massive amounts of academic audio and video content over the web, it is important to assess the performance of state-of-the-art automatic speech recognition (ASR) systems for audio/video navigation through search queries. This paper suggests a novel perspective of the challenges of ASR: instead of minimizing word error rates (WER), focus on keyword recognition. Focusing on keywords may be worthwhile for under-resourced languages, such as Hebrew, which their ASR systems have not yet reached a satisfactory accuracy level of transcription. We provide an initial Proof of Concept by demonstrating the feasible use of ASR for achieving affordable mass transcription that enables satisficing keyword recognition of a video or an audio lecture via a search engine. A forty-minutes recording set, which includes audio books and academic lectures, is used for measuring the performance of two Hebrew ASR systems, and comparing them to stenographer recordings of the video lectures, while focusing on keyword recognition. Keyness tests show advantage of keyword recognition over key-phrases results, and stenographers' records exceed both engines. Yet, keyword recognition up to 78% was achieved, which suggests that ASR has reached a satisficing accuracy level that enables its use for searching audio/video content on the web.*

**Keywords**: automatic speech recognition (ASR), audio/video search, academic video lectures, audio books, manual transcription, transcription of under-resourced languages, keyword search.

## Introduction

Face-to-face lectures have been the most common way to conduct higher education ever since the 11th century, although the printing revolution in the 15th century had shadowed this fact until the second half of the 20th century (Guri-Rosenblit, 2011). During the last five decades speech technologies have reached a certain threshold, which enables us to use them for communication purposes in general, and for teaching in particular. Indeed, nowadays, distance learning can use video lectures as a primary way of teaching (Giannakos, Chorianopoulos, Ronchetti, Szegedi, & Teasley, 2013), by providing them on the web. One of the challenges of mass video databases, which universities in general and universities that offer distance or blended learning in particular, should be thinking of, is how to gain rapid access to the contents of these video lectures, given their main conducting channel – speech.

With increasing volumes of online audio and video content, automatic transcription systems are necessary for enabling effective access by media indexation applications, which rely not only on written text, but on the audio content. Transcribing natural speech, such as academic lectures and its extreme genre – spontaneous speech – is a challenge, which Automatic Speech Recognition (ASR) research had been trying to deal with for decades (Ein-Dor, 1999; Wilpon, Rabiner, Lee, & Goldman, 1990). Therefore, we put forth some theoretical questions regarding user satisfaction. Amongst, when does accuracy become an obstacle, and a "good enough"

mechanism is just what we need, in terms of cost/effectiveness, and satisficing (Simon, 1956, 1957)? Is it necessary that a full automatic transcription will be achieved? As well as pragmatic wonders – which word error rate (WER) will be considered as acceptable? If we aim at the purpose of search queries, which words we can do without, and which are essential for the content navigation through search terms?

In this paper we present initial findings, which may be considered as a Proof of Concept (POC) for a novel perspective of the challenges of ASR – namely, not trying to do the best and to achieve the minimum error rate, but instead to be target oriented. In other words, the purpose of the present POC paper is to demonstrate the feasible use of ASR for accomplishing an idea: Achieving affordable mass transcription that enables identifying keywords of a video or an audio lecture via a search engine.

The notion of using partial information for decision-making, as opposed to complete information has been widely studied for many years in the fields of operations research (Keeney, 2009; Keeney & Raiffa, 1976), and information economics (McGuire & Radner, 1986). The main consideration in that context is cost-benefit, i.e., would it be worthwhile to pay for more information that may improve the outcomes of the decision-making process (Geri & Geri, 2011). Information economics usually regarded partial information as inferior to complete information, and the availability of more information was considered as improving the ability of the decision-maker to choose the correct course of action (McGuire & Radner, 1986). However, from the attention of economy perspective (Davenport & Beck, 2001; Geri & Geri, 2011), additional information may not always improve the decision maker performance, and may sometimes deteriorate it, as demonstrated in empirical experiments (Ahituv, Igbaria, & Sella, 1998; Geri, Neumann, Schocken, & Tobin, 2008).

In the context of this study, we are examining if partial information provided by ASR may be satisfactory for mass use of ASR transcriptions (e.g. for search purposes), and for the sake of costs reduction, enable avoiding manual transcriptions. This issue is very important for under-resourced languages, such as Hebrew, which is relatively spoken by less people than high-resourced languages, like English and Chinese. The performance of available ASR engines for under-resourced languages is usually not accurate enough for providing satisfactory full transcriptions. However, if ASR engines can transcribe most of the keywords, their output may be sufficient for enabling effective search of audio and video content.

We start with a review of the current state of audio and video search, with a focus on the role of ASR engines, as well as on Hebrew as an under-resourced language in terms of ASR technology. We then introduce our assessment method in the methodology section. We used two state-of-the-art automatic transcription systems on Hebrew audio data for the assessment process. The evaluation involved two methods: First, WER and Word Recognition Rate (WRR) measurements of the two engines were compared with the exact transcription of the audio speech. Second, we conducted keyness tests, i.e., examined whether the automatic transcription provided immediate access to main phrases (sequence of words) of the target text or whether the transcription provided immediate access to specific words (including names, people, places and organizations), mentioned in the target text. These are the keywords that would be used for searching, and therefore it would be sufficient if the ASR engines can recognize them. The results section is followed by a discussion and conclusions.

# Background and Literature Review

## Current State of Audio/Video Search

Live speech is a linear mechanism of communication. Recorded speech, as in video lectures, is not linear in that sense, since it enables back and forth navigation mechanism. However, as opposed to written texts, the 'search and find' technologies of audio databases have not yet reached a level where one can get to any point as s/he wishes. Moreover, navigation in written text is more structural, even hierarchical, since written texts consist of titles, sub-titles, paragraphs, footnotes, etc. that assist the reader. The technological "bridge" that is applied to overcome this "navigation" obstacle in recorded speech is to use written meta-data for all videos in order to provide improved access to the video contents. Nowadays, most web applications provide near-immediate access to topical lectures on the Internet, through Rich Site Summary (RSS) streams or podcasts.

Indeed, most current video search engines rely mainly on indexing the textual metadata associated with the video (title, tags, surrounding page-text). Video results that are returned by search engines are those that contain keywords in their metadata. However, sub-topics or specific content, which are not included in the metadata, are harder to locate. Even when users suspect that a certain video includes the required information, they still have to manually scan the video in order to reach their relevant content. Thus, videos, where most of the linguistic information is encoded in the audio channel, once transcribed, can be accessible as teaching materials as much as a written book, and provide topical access to automatically identified segments.

Yet, it is still far from possible for users to automatically access a certain part, within a longer audio/video lecture that might interest them. Automatic natural language recognition is considered as one of the hardest challenges of artificial intelligence (Ein-Dor, 1999). Notwithstanding, LawTo et al. (2011) claim, in the context of news broadcast videos, that contemporary advances in speech recognition systems and natural language processing have led to robust tools that enable faster, more focused access to relevant segments of one or more videos. Information retrieval (IR) from video and speech recordings is part of a concept called SEO – Search Engine Optimization (or VSEO – Video SEO). In order to increase the number of visits (i.e., traffic) of an online video, the common best practice is to provide closed captions and transcripts (Dugdale, 2010). As long as the lecture is transcribed, or a video has captions, the navigation obstacle can disappear, since the speech channel is then transferred into a textual medium. The automatic closed caption of YouTube's videos is only one example (Robertson, 2010), where automatic transformation of captions into raw text file (text-to-text transformation), with time-stamps, is also available. Raw text file transcriptions are also available in Ted-Talk videos, but with no time-alignment. Concerning Hebrew, both sites provide Hebrew transcripts, although it is unclear whether the transcripts are direct from ASR's engine or a translation of the transcripts in the source language.

Transcription of audiovisual lectures is essential for learning processes since it serves as anchors within a flood of linear information (speech). With transcription, the user can easily skip from one point to the other. Indeed, such search capability can increase the usability of videos, and improve learning, since students will get just what they need – Focusing. This capability can help

students create their own texts and notes, and integrate them with the lecture, exactly as they add notes on digital books. In Ronen, Raz, & Akam (2014), a college teacher was asked to give feedback in two different ways: written feedback and recorded video feedback. The main limitation that the teacher reported was the need to replay on and on the video feedback in order to search for specific information, information that in the written feedback was quickly located. Moreover, it was found that 83% of the students summarized the highlights of the recorded video feedbacks compared to 25% of students who summarized the written feedbacks. This research demonstrated the actual necessity of transcribed video learning materials (lectures, feedbacks, etc.). It also indicated that in certain genres, the transcriptions should be precise. This creative and active task is achievable by existing software (e.g., SubPly.com, ramp.com). Some of these tools provide a time stamp of each note, so when applied, the video is aligned with certain amount of text, depending on the student's activity, just like captions provide orientation. There are few tools available to provide transcription and closed captions for video files. For example, SubPLY.com provides free automatic transcription and closed captioning for English videos. Also, one can take such transcripts and upload them to a YouTube account as YouTube will then use Google's speech-to-text recognition technologies to transform plain text transcript into closed captions for YouTube videos.

To understand the challenge of video and audio transcription means to acknowledge that, currently, full access to speech data and flexible navigation in speech data can be achieved only by speech-to-text transformation. The question is whether this transcription should be performed manually, automatically, or jointly.

## Automatic Speech Recognition

Automatic Speech Recognition is a technology that aims to recognize human speech and transcribe it. Such systems are already in use in highly-resourced languages, such as American-English and Japanese. This is to say that Speech Recognition (SR) technology is language dependent, and in order to apply it to a new language, the system has to be first trained on it. To date, most of SR advancements have been in other languages, primarily English. Hebrew, on the other hand, is considered an under-resourced language, since there is no open resource of SR infrastructure, mainly, a large speech database, to allow the development of Hebrew SR.

An under-resourced language means that the linguistics and technological infrastructure does not exist for common use. Moreover, a simple search on the online catalogue provided by LDC – The Linguistic Data Consortium (the global largest distribution agency for language resources), illustrates the situation: LDC is an open consortium for universities, companies and government research laboratories. As of January 2014, among LDC's registered corpora, 363 were for English, 16 for French, 85 for Mandarin Chinese, and 116 for Arabic. A search for Hebrew language resources was fruitless. It suggests that the business sector does not view such an infrastructure for the Hebrew language as having a significant potential for the industry. On the other hand, the research community has difficulty funding such an activity. A step forward toward a Modern Hebrew state-of-the-art infrastructure (a speech database and its phonetic transcription) was reported in Silber-Varod, Latin, and Moyal (2013).

Automatic Speech-to-Text technology varies in terms of the recognized speech material. An overview of Word Error Rate in various ASR tests compared to human perception was described in Lippmann (1997). Lippmann (1997) demonstrates how WER escalates the more complex the speech material is. From less than one percent human versus machine error rates of 0.105% and 0.72%, respectively, for a digit corpus, to human Error Rate (ER) for continuously spoken letter and machine ER for isolated spoken letters (1.65% and 5%, respectively). Lippmann (1997) then shows how null grammar sentences (i.e., sentences where a machine recognizer assigns equal probability to all words independent of the surrounding words, which gives an effect of understanding nonsense sentences by human), and larger vocabulary of 1,000 words, escalate WER, both for human perception and machine transcription (2% and 17%, respectively). WER is also largely affected by non-linguistic conditions, such as noisy environment, variety of recording environments, sound effects and multiple speakers. Lippmann (1997) shows how machine ERs are roughly ten times higher than those of humans, and increase without noise compensation (increase at the lower Speech-to-Noise Ratios). Nonetheless, large vocabulary continuous speech recognition (LVCSR) achieved the worst machine WER results – 43%, compared to 4% of human WER (Figure 1). Another significant parameter for ASR performance is speech styles: In an experiment reported in Lippmann (1997), talkers engaged in spontaneous conversations were asked to read the transcriptions of these conversations. The machine recognition error rate was 52.6% for spontaneous speech and 28.8% for read versions of the same materials. Rousseau, Deléglise, and Estève (2012) have built an ASR based on the TED Talks leading to a WER score of 17.4%.

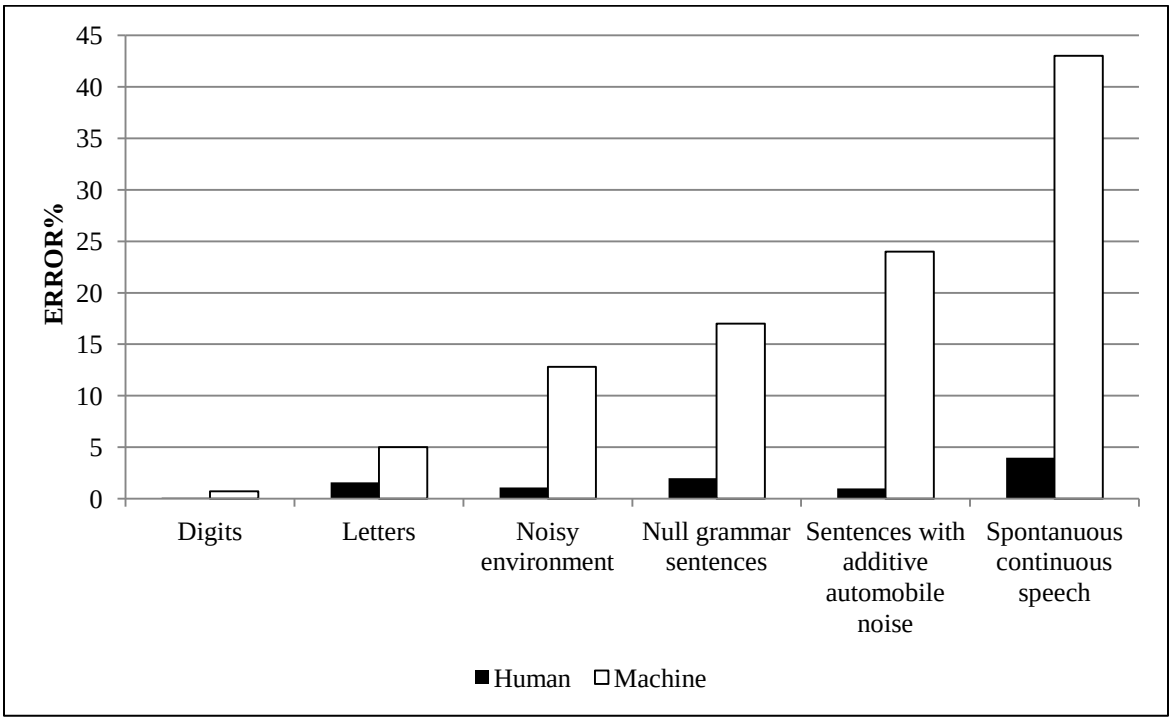



**Figure 1: Human versus Machine error rates for the various types of speech corpora (summary after Lippmann, 1997)**

A comparison between automatic audio and human manual transcripts alignment was carried by Barras et al. (2004) for French radio speech. They have shown that a combined automatic and manual system, where the human annotator could concentrate on specific "problematic" subpart of the signal, performed about eight times real-time (i.e., eight times the actual duration of the speech segment), while the estimated manual post-processing correction time reported in Lamel and Gauvain, (2003) is 5XRT (i.e., five times real-time) for closed captions of broadcast news in American English. Both reported results should be compared with the 60 times real-time manual production of a precise orthographic transcription synchronized with the audio signal, when no a-priori information is provided (Barras et al., 2004). Moreover, transferring time-codes to the written documents "may be helpful in an audio archive or audio information retrieval environment." (Barras et al., 2004). The reported ASR in Barras et al. (2004) exhibited an average WER of 12%. About 80% of the text (with word chunks of at least five words) could be exactly aligned with the automatic transcripts of the audio data. The residual WER on these 80% was less than 1%.

Every ASR engine relies on linguistic infrastructure that is crucial to the ASR output. Word level language model, for example, is a mechanism by which the engine is conducting post-processing (i.e., post signal processing) of the recognized words, by taking word sequence probabilities in a specific language into consideration (for example, the sequence "in the" is more probable than "the in"). Thadani, Biadsy, and Bikel (2012) hypothesized that decoding the speech of a particular video using a language model adapted to the topics of that video would improve automatic transcription. This technique assigns topics to utterances using N-best recognition hypotheses produced under generic models, and then obtains a final transcription under a topic-adapted model. Thadani et al. (2012) observed that the inclusion of topics (i.e., context) significantly improved WER. For Google's ASR, WER was reported as 24.8% for cross-dialect Arabic voice search (Biadsy, Moreno, & Jansche, 2012). In the Arabic dialect research, users spoke their search queries, typically using a mobile phone, and the system returned a transcription and web search results.

Our research seeks to investigate how we can significantly reduce the gap between machine and human performance for the purpose of Hebrew text navigation through search terms. The purpose of this study is therefore to examine rapid and affordable ways to transcribe Hebrew speech, by existing tools, and to explore their potential to provide good enough, not perfect though, video transcriptions. As technology-based tools suggest long-term cost reduction, our proposal is a step forward toward such automaticity.

An assessment of automatic transcription with regard to its written transcriptions version can be perceptual, by asking users if they find the transcriptions satisfying for their purposes. Eklund (2012), for example, relies on hundreds of quality assurance (QA) reports of commercial applications, and "whether or not the reported *Successful* and *Ignored* rates were produced by the same or different labelers" (Eklund, 2012). Another method is to use objective values of WER, by comparing the automatic transcriptions to the reference original transcriptions. Assessment is also achieved by comparing human recognition to the ASR's (e.g., Lippmann, 1997), Yet, when aiming at search queries purposes, the two versions should be compared not only according to quantity parameters of WER, but also according to quality parameters, that consider the recognized word types, or word semantic significance. i.e., keywords.

## Research Hypotheses

Given the theoretical background, we hypothesize the following:

> *H1: Audio book WRR of ASR would be better (i.e., smaller) than that of public lectures.*
>
> *H2: Exact offline manual transcriptions would show a better match to ASR output than a stenographer's (online manual) records.*

H1is expected due to the recording conditions: Recordings of read speech in an acoustic room versus recordings in a crowded auditorium. H2 assumes that since real-time stenographers know that the aim of their work is accessibility of hearing impaired, and not a legal document (such those recorded by a court reporter), then they put less efforts on accuracy and more on delivering the main ideas.

Keywords are considered as content words, which are more likely to appear in the linguistic infrastructure of the ASR engines (lexicon, and in their language model). Therefore, we suggest the following:

> *H3: Keyword recognition rate will be higher than the general word recognition rate (of the same engine).*

# Methodology

The performance of machines and humans is compared in this paper using word recognition rate, which is derived directly from word error rate, a common metric of the performance of speech recognition. The analysis includes the following parameters: recording type, speech genre, and gender differences, and the keyword results discussion is conducted with linguistic terminology.

## Data and Transcriptions

The database consisted of about 40 minutes of Hebrew speech. Forty minutes are considered a small size corpus for ASR experiments. For example, Google training set for their Hebrew engine was 265K searches per training, and 15K searches for testing (Biadsy, 2013). This is estimated by 933 hours, since each search is maximum 12 seconds long. Yet, Google's purposes are different from our goal in the sense that we investigate continuous speech recognition, not short, single spoken word/phrase search queries, and our purpose is to examine if the output contains certain keywords of the original audio file. In Thadani et al. (2012) 200 hours were trained. One hour to 62 hours is reported in Lippmann (1997) for various ASR experiments, all of which were experimenting novel algorithms for ASR engines. Our goal, on the other hand, was not to prove ASR or to suggest a new algorithm, but to test the possibility and to provide proof-of-concept documentation of using off-the-shelf engines. Table 1 presents the scheme of the four-parts database.

As to the recorded material, the two academic lectures were carried during a conference in July, 2013. Both were technology oriented and discussed different aspects of accessibility on the internet. The first audio book is an Organizational Behavior textbook, recorded by a professional

male narrator; the other is a Network Economy textbook recorded by a professional female narrator. All recording were carried at the Open University of Israel campus.

**Table 1: The database scheme**

| | **Audio book** | | **Academic lecture** | | **Total** |
|---|---|---|---|---|---|
| **Speakers** | Woman | Man | Woman | Man | |
| **Duration (minutes)** | 10 | 10 | 10 | 10 | 40 |
| **# of words** | 1,038 | 1,005 | 991 | 1,048 | 4,082 |
| **# of audio fragments (12 seconds each) for the Google engine** | 50 | 50 | 50 | 50 | 200 |
| **# of audio fragments (1 minute each) for the NDEV engine** | 10 | 10 | 10 | 10 | 40 |

## ASR Engines

The two state-of-the-art automatic speech recognition engines that were used are: Google/HTML5 speech recognition system for Hebrew (Sampath & Bringert, 2010) and Nuance Mobile Developer Program - NDEV (2011). Both are closed tools with no possibility to change their acoustic models, and linguistic infrastructure: lexicon (i.e., word list and transcriptions) or Language Model. Google Voice Search is a free access engine with an Application Program Interface (API). It enables a single query of an audio file (12 seconds long, FLAC format) as an input, and turns back results in JSON format as an output, which is translated into textual format (i.e., transcription). NDEV Mobile is Nuance free product that enables flexible access to their speech models and SR engine. It also has an API interface. The audio file required format for both engines is mono *.Wav files, with 16kHz sampling rate, 16bit PCM. NDEV audio files were up to 1 minute long. The recognition tests were carried during December 2013, and are therefore relevant to the engines' versions at that time. These engines are said to be constantly updated and recognition rates may be changed with each version.

## Performance Measures

For each test, the WER is calculated in comparison to a manual transcription reference. WER is derived from Levenshtein Distance Measure that is calculated at the word level and is used to measure the difference between two sequences in information theory. The WER is calculated according to the following formula:

$$\mathrm{WER} = \frac{S + D + I}{N}$$

Where:

$S$ = number of substitutions. Substitution: A word in the automatic transcriptions that is aligned to a non-identical word in the corresponding manual transcription.

$D$ = number of deletions. Deletion: A word in the manual transcriptions that is not aligned to any word in the corresponding automatic transcription.

*I* = number of insertions. Insertion: A word in the automatic transcription that is not aligned to any word in the corresponding manual transcription.

*C* = number of correct scores. Correctness: A word in the automatic transcription that is aligned to an identical word in the corresponding manual transcription.

*N* = number of words as an input: C+D+S = N

*WER* can also be reversed as Word Recognition Rate (WRR): 1− WER = WRR.

The reference written material consisted of three different types of written texts. The first reference of the audio books was simply the parallel sections in the textbooks. As for the academic lectures, there were two references in two tests. A. Stenographer's records, which were carried on real time, during the lectures themselves; B. Exact transcriptions that were carried manually. The exact transcriptions contain all speech events that were uttered by the lecturers, including lexical words, but also hesitations, repetitions, and false-starts. All punctuations marks were deleted from all three types of reference texts.

## Keyness

*Keyness* is a term used in linguistics to describe the quality a word or phrase has of being "key" in its context. Keyness is a *textual* feature, not a language feature (so a word has keyness in a certain textual context but may well not have keyness in other contexts. In the keyness tests, an expert was asked to mark the keywords or key-phrases (i.e., a sequence of adjacent words) in the four corpora. The expert was told to think about the semantically significant words or the important names, phrases and content words in each of the exact texts. The goal of the keyness test was to measure whether the ASRs output texts provide immediate access via search engines to main topics of the *exact* transcriptions; and whether the text provides immediate access to events, people, places and organizations, mentioned in the *exact* transcriptions.

This last point is important since contemporary search engines do not necessarily need the exact form of a word and search engines technology knows how to cope with different forms of the same "basic" word (i.e., *lemma*, which can be defined as the dictionary form of a word),. This "restriction" of matching only the *exact* string and not words with affixes (e.g., *organization* and not organizatio*nal*) or with clitics (e.g., organization vs. organizations) is a methodological choice, to enable comparison to tests A-C, which were carried with the *exact* transcriptions as a reference, and which rely on exact match. To sum up the Keyness experiment, practically, the test was carried in order to see if the ASR output is sufficient as an input for an effective accurate search process, although the ASR does not provide complete accurate results.

## Tests

Following the above methodological outline, the following tests were carried:

A. WRR: Automatic transcriptions of the two engines versus exact (manual) transcriptions.
B. WRR: Automatic transcriptions of the two engines versus stenographers' manual records.
C. Textual WRR: Stenographers' manual records versus exact manual transcriptions.

D. Key-phrases and keywords: ASR of the two engines versus an expert's manual annotation.
E. Key-phrases and keywords: Stenographers' records versus expert's manual annotation.

## Results

The WRR results in tests A-C are presented first, followed by keyness results in tests D-E. Figure 2 summarizes the WRRs of the two ASR engines in tests A-B, and it demonstrates the expected increase of WRR performance between the lecture tests (LM denotes a lecture by a man and LW denotes a lecture by a woman) and the book tests (BM denotes a book read by a man and BW denotes a book read by a woman), which supports H1. The WRR of the two engines versus stenographers' records in the two lectures are represented in Figure 3. A considerable gap between man and woman recognition rate is demonstrated, as well as difference in Lecture man's WRR between test A and test B.

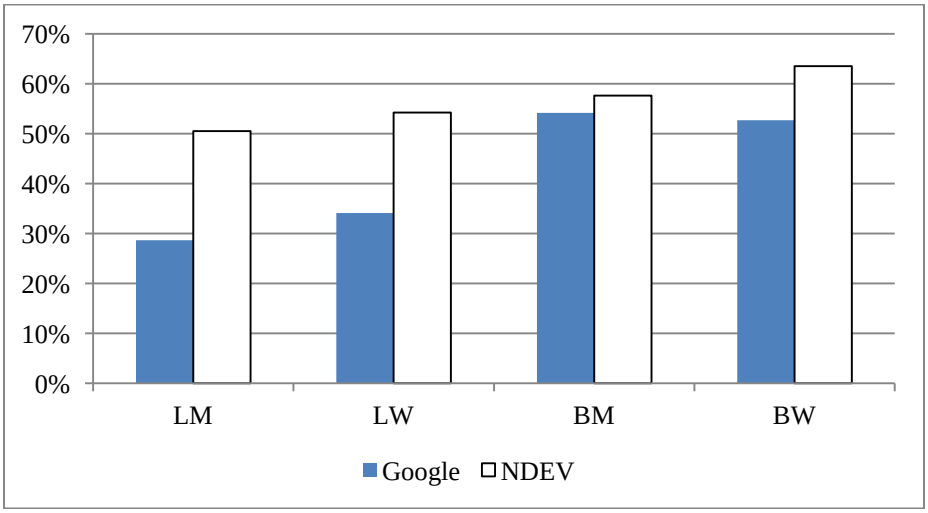



**Figure 2: Comparison of WRRs with regards to recording type (test A)**

Following the difference between the two tests, A and B, we conducted a comparison, in WER terms, of the stenographer's records to the exact transcriptions, with the latter as a reference (Test C). The first observed difference is the word count. While the exact transcriptions consisted of 1000 words per lecture, the parallel text of the stenographer's records consisted of only 847 words in LW, and 744 words in LM. This reduction of words predicts the higher amount of deletion rates. Indeed, as shown in Figure 4, deletions are the main cause of the WERs in both lectures (36.93%, which is 57% of WER in LM, and 16.04%, which is 52% of WER in LW), but substitutions as well (31% and 43% of WER, respectively). Although these two lectures were carried at the same day and recorded by the same stenographer, WRR of LW is twice larger than WRR of LM. As to H2 – results of test C suggest that real-time stenographers would put less effort on accuracy and more on bringing the main ideas of the lecture.

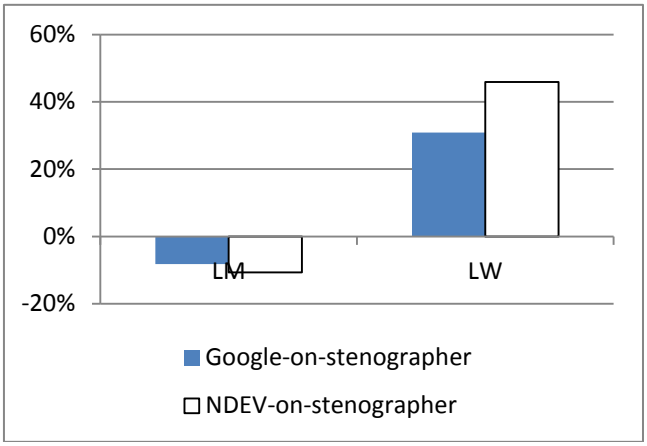
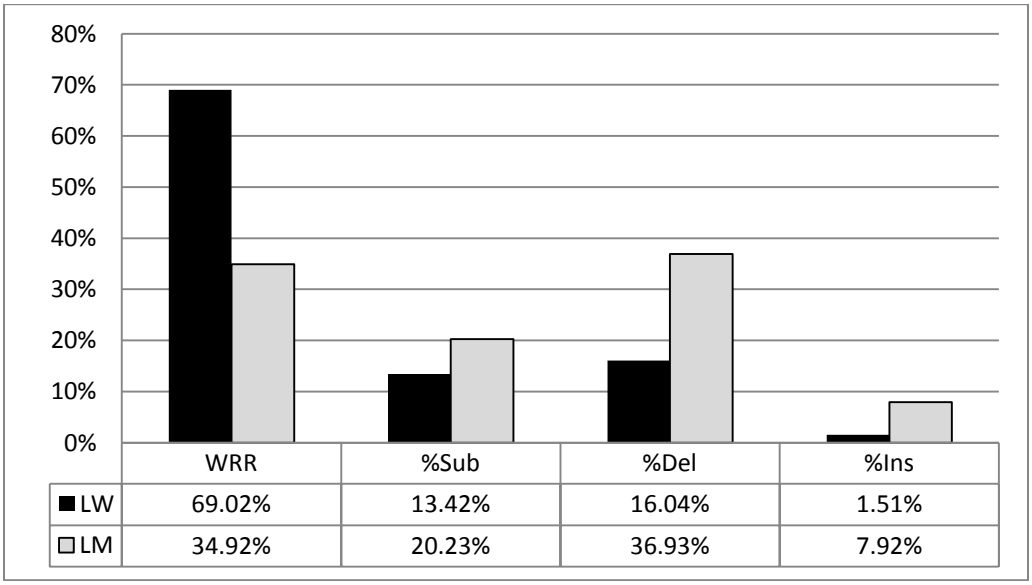



**Figure 3: Comparison of WRRs on the stenographers' records in the two lectures (test B)**




| | WRR | %Sub | %Del | %Ins |
|---|---|---|---|---|
| LW | 69.02% | 13.42% | 16.04% | 1.51% |
| LM | 34.92% | 20.23% | 36.93% | 7.92% |

**Figure 4: A comparison of stenographer's records and the exact transcriptions**

The results of tests D-E are demonstrated in Tables 2-5, and Figure 5. In Table 2 the key-phrase results of Book-Man are presented (right columns), where it is demonstrated that about 50% of key-phrases (phrases are considered having between 1 to 5 words) were detected by the ASR engines. Because phrases are statistically less probable to be recognized since they are a matter of language model and less of acoustic models, a keyword test was carried, meaning, a single word list was extracted from the expert key-phrases list (for example, [hitnahagut] 'behavior' and [irgunit] 'organizational' are two different words in the list. The results are demonstrated in Table 2 (left columns), where it is shown that both ASRs do better recognition on a word-level basis. In

Tables 3-5 the keyword and key-phrase recognition results of the three other corpora are presented.

**Table 2: Keyword and key-phrase recognition rates in Book-Man**

| | **Keyword recognition** | | | | **Key-phrase recognition** | | | |
|---|---|---|---|---|---|---|---|---|
| | # types | #occurrences | % types | %occurrences | # types | #occurrences | % types | %occurrences |
| **Exact** | 170 | 411 | 100% | 100% | 119 | 198 | 100% | 100% |
| **Google** | 125 | 291 | 74% | 71% | 62 | 132 | 52% | 67% |
| **NDEV** | 124 | 335 | 73% | 82% | 65 | 115 | 54% | 58% |

**Table 3: Keyword and key-phrase recognition rates in Book-Woman**

| | **Keyword recognition** | | | | **Key-phrase recognition** | | | |
|---|---|---|---|---|---|---|---|---|
| | #types | #occurrences | %types | %occurrences | # types | #occurrences | % types | %occurrences |
| **Exact** | 139 | 343 | 100% | 100% | 101 | 172 | 100% | 100% |
| **Google** | 104 | 250 | 75% | 73% | 59 | 104 | 58% | 60% |
| **NDEV** | 94 | 273 | 68% | 80% | 56 | 127 | 55% | 74% |

**Table 4: Keyword and key-phrase recognition rates in Lecture-Woman**

| | **Keyword recognition** | | | | **Key-phrase recognition** | | | |
|---|---|---|---|---|---|---|---|---|
| | # types | #occurrences | %types | %occurrences | # types | #occurrences | % types | %occurrences |
| **Exact** | 130 | 305 | 100% | 100% | 87 | 125 | 100% | 100% |
| **Stenog.** | 111 | 254 | 85% | 83% | 67 | 100 | 77% | 80% |
| **Google** | 87 | 239 | 67% | 78% | 41 | 60 | 47% | 60% |
| **NDEV** | 87 | 224 | 67% | 73% | 42 | 64 | 48% | 51% |

**Table 5: Keyword and key-phrase recognition rates in Lecture-Man**

| | **Keyword recognition** | | | | **Key-phrase recognition** | | | |
|---|---|---|---|---|---|---|---|---|
| | # types | #occurrences | %types | %occurrences | # types | #occurrences | % types | %occurrences |
| **Exact** | 42 | 88 | 100% | 100% | 30 | 45 | 100% | 100% |
| **Stenog.** | 29 | 65 | 69% | 74% | 18 | 29 | 60% | 64% |
| **Google** | 18 | 68 | 43% | 77% | 11 | 21 | 37% | 47% |
| **NDEV** | 19 | 58 | 45% | 66% | 13 | 21 | 43% | 47% |

In comparison to the *general* word recognition rates (which are presented in Figure 2), Figure 5 demonstrates higher Keyword and key-phrase recognition rates. These findings support H3. The

*general* WRR, shown in Figure 2, range from 28.63% (in Lecture-Man) to 63.49% (in Book-Woman), with an average recognition rate of 49.42%. The *keyword* recognition rates, on the other hand, range from 66% (NDEV LM, top black line in Figure 5) to 82% (NDEV BM, top black line in Figure 5), with an average of 76%; The *Key-phrase* recognition rates range from 47% (both Google and NDEV LM, bottom black line in Figure 5) to 74% (NDEV BW, bottom black line in Figure 5), with an average of 61%. It is worth noting that the mentioned figures concern occurrences ratios, not type ratios, which, as demonstrated in the two gray dotted lines, have lower WRR (average of 67% for keyword types and of 53% for key-phrase types).

As to internal comparison of keyword and key-phrase WRRs, It is evident that read books have better results than lectures. No gender differences were detected as the book tests show better results of man recognition while lecture tests demonstrate better results of woman's recognition. Stenographer's records demonstrate the best keyword results, which is not surprising since these records are filtered on the spot in order to present main issues that have been said in the lectures.

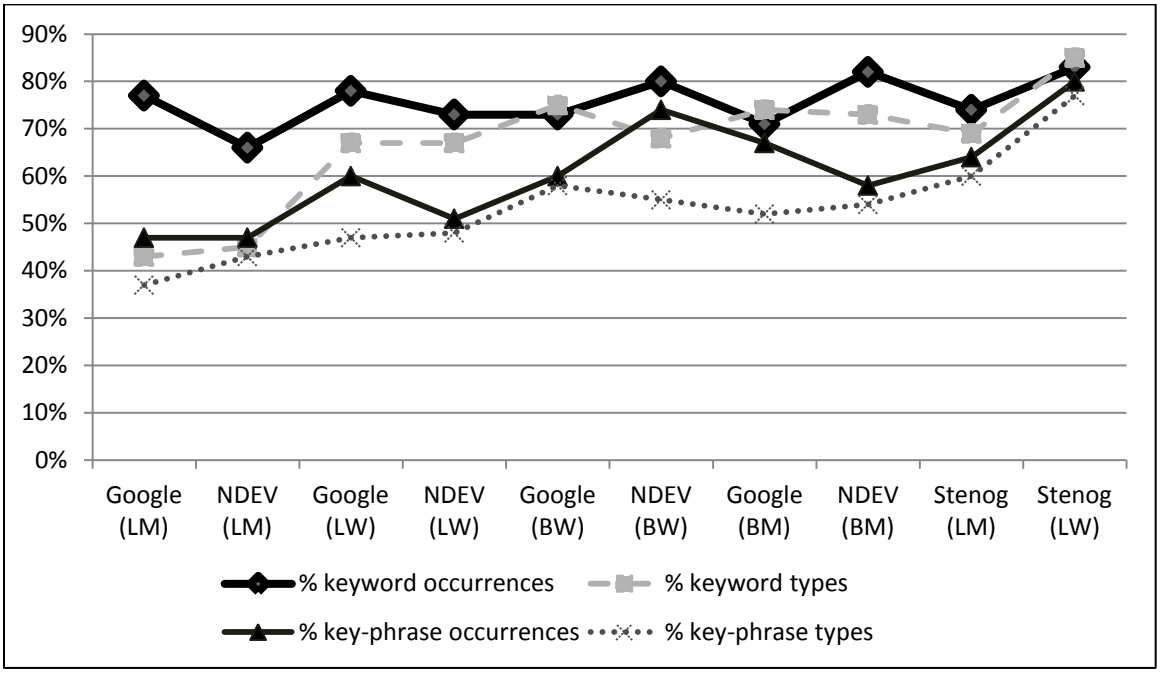



**Figure 5: Comparison of recognition rates of keywords and key-phrases (type and occurrences) of the four recordings: Book-Man (BM), Book-Woman (BW), Lecture-Man (LM), and Lecture-Woman (LW)**

## Discussion

## Theoretical Implications

As H1 proposed, the WRR tests showed that the ASRs performed better with read speech than with lectures, in quantity and quality. We then compared two transcriptions: Exact and stenographer's records. The engines' automatic transcriptions were closer to the exact transcriptions than to the stenographer's records, so H2 was also supported by our findings. This

comparison means that stenographer's records, which are the standard way to have on-the-spot transcriptions during academic lectures (the stenographer's records are presented on a large screen beside the lecturers), provide 69.02% word *recognition* rate, in the better case of Lecture-Woman, and 34.92%, in the worse case (of Lecture-Man), similar to the ASR performance in almost all parts of test A: the book results, Lecture-Woman results, and Lecture-Man NDEV results.

As to the keyness tests, H3 was supported by our findings, which showed that keyword recognition rate was higher than the general WRR of the same engine. Stenographers' records were almost the same as NDEV results of Book-woman and Book-man which reached 80% keyword recognition rate. Yet, as explained above, the key-phrase or keyword queries should be conducted with a standard search engine in order to provide *lemma* results and not *exact spelling* results. This is especially crucial when searching for key-phrases. For example, the "exact match" method, which was adopted in the present research, found the word [atar] 'site' in Lecture-Man, three times, exactly as it was in the original text, but this lemma-word was also repeated in two other forms and in one phrase, which were not found, while a simple search query of the lemma [atar] 'site' retrieved all the four forms (the lemma and the three other forms). Gender differences cannot be explained by technological reasons, since ASR engines are known to be trained by a balanced training set, with equal part to men speech and women speech (this does not mean that the speech signal has no gender characteristics. on the contrary (Cieri, Miller, & Walker, 2004)). The explanation of the difference between Lecture-Woman and Lecture-Man should probably be due to one of the followings: personal differences, their voice quality, their use of foreign words, or even due to lack of awareness on behalf of the stenographer during Lecture-Man.

As for the theoretical question whether partial information may be satisficing for certain purposes, in the context of this study we demonstrated how partial information provided by ASR may be satisfactory for mass use of ASR transcriptions (e.g. for search purposes), and for the sake of costs reduction, enable avoiding the use of manual transcriptions. The performance of available ASR engines for under-resourced languages, such as Hebrew, is usually not accurate enough for providing satisfactory full transcriptions. However, since we provided initial evidence that ASR engines can transcribe most of the keywords, their output may be sufficient for enabling effective search of audio and video content.

## Practical Implications

Practical implications of the present research are threefold. *Learning aspects*: Transcription of audiovisual lectures is essential for learning processes since it can serve as anchors for navigation within the video. With transcription, the flood turns into a non-linear structure, like any other written text. Indeed, such search capability can increase the use of videos, and improve learning skills; *Accessibility*: Additionally and arguably more important, using transcriptions and closed captions increases accessibility. Accessible format of academic content will serve the hearing-impaired, but also a considerable proportion of the elderly population that suffers from gradual hearing decline; *Cost/benefit aspects*: Although speech-to-text recognition technologies at this point are nowhere near 100% accurate, initial transcriptions can be generated automatically for any audio document via automatic speech recognition, and can be used for

search queries of keywords, and for navigation within the video. Another use of the initial transcriptions can be as a draft for manual transcription, and thus reduce process rate and make it closer to real-time. The quality of the automatic transcript depends on the ASR.

## Limitations and Further Study

In future research we intend to also use subjective quality assessment, and to question Hebrew speakers about their subjective assessment of the transcription and search capability. As to language-dependent aspects, every ASR engine relies on linguistic infrastructure that is crucial to the ASR output. From our tests it became evident that a Word level language model must include technical words and sometimes even to adapt code-switching (i.e., alternating between two or more languages) models, since natural speech of academic lecturers often includes such characteristics, especially in technological and innovation issues, and in textbooks as well.

## Conclusions

This paper has provided an initial proof of concept that ASR engines may provide satisficing transcription that would enable effective search of video and audio content. Our empirical tests of Hebrew ASR demonstrated nearly 80% recognition of keywords. The concept of focusing on keywords should be primarily important for under-resourced languages, which their ASR systems have not reached yet a satisfactory accuracy level of transcription. By demonstrating the feasibility of using an ASR for achieving affordable satisficing mass transcription, this research contributes a valuable insight for improving ubiquitous global consumption and management of knowledge.

## Acknowledgement

The authors gratefully acknowledge that this research was supported by the Open University of Israel's research fund (grant no. 502532).

## References

Ahituv, N., Igbaria, M., & Sella, A. (1998). The effects of time pressure and completeness of information on decision-making. *Journal of Management Information Systems, 15*(2), 153-172.

Barras, C., Adda, G., Adda-Decker, M., Habert, B., Boula de Mareuil, P., & Paroubek, P. (2004). Automatic audio and manual transcript alignment, time-code transfer and selection of exact transcripts. *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC'04)* (pp. 877-880). Lisbon.

Biadsy F. (2013). Google's voice search – focusing on Arabic and Hebrew. Presentation at *ISCOL 2013 conference* (June 2013), Ben Gurion University. http://www.cs.bgu.ac.il/~cohenrap/iscol2013/program.html#abstracts

Biadsy, F., Moreno, P. J., & Jansche, M. (2012). Google's cross-dialect Arabic voice search. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)* (pp. 4441-4444).

Cieri, C., Miller, D., & Walker, K. (2004, May). The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In *LREC* (Vol. 4, pp. 69-71).

Davenport, T. H., & Beck, J. C. (2001). *The attention economy: Understanding the new currency of businesses*. Boston, MA: Harvard Business School Press.

Dugdale, D. (2010). The Online Video Marketing Guide. A video available at: http://www.reelseo.com/transcribing-videos-automated-transcription/

Ein-Dor, P. (1999). Artificial intelligence: A short history and the next forty years. In K. E. Kendall (Ed.), *Emerging information technologies: Improving decisions, cooperation, and infrastructure* (pp.117-140). Thousand Oaks, CA: Sage.

Eklund, R. (2012). ASR "sweet sixteen": An evaluation of Nuance Swedish speech recognizer success rates in 69 commercial applications 16 years after its inception and an assessment of inter- and intralabeler agreement. *Proceedings of FONETIK 2012* (pp. 113–116). Gothenburg, Sweden, May 30–June 1, 2012.

Geri, N., & Geri, Y. (2011). The information age measurement paradox: Collecting too much data. *Informing Science Journal*, *14*, 47-59.

Geri, N., Neumann, S., Schocken, R., & Tobin, Y. (2008). An attention economy perspective on the effectiveness of incomplete information. *Informing Science Journal*, *11*, 1-15.

Giannakos, M. N., Chorianopoulos, K., Ronchetti, M., Szegedi, P., & Teasley, S. D. (2013). Analytics on video-based learning. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 283-284). ACM.

Guri-Rosenblit, S. (2011). Universities: Moving from a national system to a glocal network. In I. Tubella & B. Gros (Eds.), *Turning universities upside down: Actions for the near future* (pp. 151-176). Barcelona: Universitat Oberta de Catalunya Press.

Keeney, R. L. (2009). *Value-focused thinking: A path to creative decisionmaking*. Boston, MA: Harvard University Press.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: preferences and value trade-offs*. New-York: Wiley.

Lamel, L. & Gauvain, J. L. (2003). Linguistic data: Fast transcription. *The rich Transcription Spring 2003 Evaluation (RT-03S) workshop*. Boston, MA.

LawTo, J., Gauvain,J. L., Lamel, L., Grefenstette, G., Gravier, G., Despres, J., Guinaudeau, C., Sebillot, P. (2011). A scalable video search engine based on audio content indexing and topic segmentation. *Proceedings of the Networked and Electronic Media (NEM) Summit: Implementing Future Media Internet*.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, *22*, 1-15.

McGuire, C. B., & Radner, R. (Eds.). (1986). *Decision and organization* (2nd ed.). Minneapolis, MN: University of Minnesota Press.

Nuance Mobile Developer Program (2011). HTTP Services for Nuance Mobile Developer Program Clients. Nuance Communications Inc. Retrieved from http://dragonmobile.nuancemobiledeveloper.com/public/Help/HttpInterface/HTTP_Services_for_NMDP.pdf

Robertson, M. R. (2010). In-Depth Look at YouTube Closed Captions - YouTube SEO and More. Retrieved from: http://www.reelseo.com/youtube-closed-captions-seo/

Ronen, M., Raz, R., & Akam, S. (2014). Screencast feedback to students' artifacts: Potential & challenges. In Y. Eshet-Alkalai, A. Caspi, N. Geri, Y. Kalman, V. Silber-Varod, Y. Yair (Eds.), *Proceedings of the Chais conference for Innovation and Learning Technologies 2014: Learning in the technological era* (pp. H183-H192). Raanana: The Open University of Israel [in Hebrew].

Rousseau, A., Deléglise, P., & Estève, Y. (2012). TED-LIUM: an automatic speech recognition dedicated corpus. *Proceedings of LREC 2012, the Eighth International Conference on Language Resources and Evaluation* (pp. 125-129). Istanbul (Turkey), May 21-27, 2012.

Sampath, S., & Bringert, B. (2010). Speech Input API specification, Google Inc. W3C 2010. Available at: http://lists.w3.org/Archives/Public/public-xg-htmlspeech/2011Feb/att-0020/api-draft.html

Silber-Varod, V., Latin, M., & Moyal, A. (2013). Hebrew phoneme database: A project that begins with 100.000.000 Words and ends with 31 phonemes. A paper presented at *The 29th Annual Meeting of the Israeli Linguistics Society*. February 4, 2013, Sefad, Israel. [Hebrew].

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129.

Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley.

Thadani, K., Biadsy, F., & Bikel, D. (2012). On-the-fly topic adaptation for youtube video transcription. *Proceedings of INTERSPEECH*, Portland, USA, September 2012.

Wilpon, J. G., Rabiner, L. R., Lee, C., & Goldman, E. R. (1990). Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models. *IEEE Transactions on Acoustic Speech Signal Processing*, *38*(11), 1870-1878.

## Biographies

**Vered Silber-Varod** is a Research Fellow at The Research Center for the Study of Innovation in Learning Technologies, The Open University of Israel. She holds a B.A. in Political Science and French language and literature studies from the Hebrew University of Jerusalem, an M.A. with Magna Com Lauda in Hebrew language studies at the department of Hebrew and Semitic languages from Tel Aviv University, and a Ph.D. in Humanities from Tel Aviv University. Vered has over a 10 years exprience as a Linguist at various institutes, including the ACLP – Afeka

Center for Language Processing, Afeka College of Engineering. Her research interests and publications focus on various aspects of linguistics, with expertise in speech prosody, acoustic phonetics, and quantitative analysis of written and spoken texts. Dr. Silber-Varod is a member of the International Speech Communication Association (ISCA). Personal site: http://www.openu.ac.il/Personal_sites/vered-silber-varod/

**Nitza Geri** is a faculty member at the Open University of Israel, Department of Management and Economics, and Head of the Research Center for Innovation in Learning Technologies. She holds a B.A. in Accounting and Economics, an M.Sc. in Management Sciences and a Ph.D. in Technology and Information Systems Management from Tel-Aviv University. Nitza is a CPA (Israel), and prior to her academic career she had over 12 years of business experience. Her research interests and publications focus on various aspects of the value of information, and information systems adoption and implementation, including strategic information systems, e-business, economics of information goods, attention economy, knowledge management, value creation and the Theory of Constraints, managerial aspects of e-learning systems adoption and use. Personal site: http://www.openu.ac.il/Personal_sites/nitza-geri.html