Multimodal speech recognition errors and second language acquisition: An activity theoretic account

Rod Farmer
raf@cs.mu.oz.au
Department of Computer Science and Software Engineering
The University of Melbourne

Abstract Researchers in Computer-Aided Language Learning (CALL) have long recognized the potential benefits of Human Language Technology (HLT) for affording more naturalistic and authentic language learning contexts. Historically, CALL practitioners have argued that advancements in HLT would provide more intuitive, adaptive and compelling human-computer interfaces for effective language learning environments. However, CALL researchers have yet to demonstrate an understanding of how multimodal interaction affects the language learning process. Moreover, there are few Human-Computer Interaction (HCI) studies in CALL that critically examine the relationship between multimodal interaction, cognitive complexity and language acquisition.

HCI methodologies in CALL have traditionally been couched within the cognitivist (information processing) paradigm. However, there is growing awareness amongst HCI practitioners of the importance of post-cognitivist, situated methodologies in modeling social systems, such as language learning. In contrast to cognitivist approaches, the goal of socially-oriented methodologies is to uncover emerging themes or patterns across subjective interpretations of user experiences. Such methodologies are therefore more relativistic and qualitative in nature. It is argued here that post-cognitivist HCI frameworks are more amenable to capturing the complexities of learner-computer interaction in CALL.

This paper submits two contributions to the study of language learning and technology. Firstly, the paper presents initial findings from an exploratory study that undertook the development and evaluation of multimodal speech technology for second language vocabulary acquisition. Secondly, an activity theoretic examination of these findings is proffered. Findings from this study suggest future research agendas for the design and evaluation of multimodal systems in learner-centred environments.

1 Introduction

Researchers in Computer-Assisted Language Learning (CALL) have witnessed over the past decade rapid advancements in the field of Spoken Language Technology. Hardware advances (processor speed, storage, size) and theoretical contributions from the fields of natural language processing, machine learning and digital signal processing have greatly contributed to the successful commercialisation of many speech systems. In contrast to the speech systems of a decade ago, most commercially available speech platforms today support large vocabulary, open-continuous automatic speech recognition (ASR). Additionally, Text-to-Speech (TTS) systems have reached the point where, in some domains of use, it is difficult to differentiate between synthetic and natural language. Although there is undoubted need for continued efforts to improve the robustness, naturalness and intelligence of spoken language interfaces, many believe that the major challenges of last century have been solved. Indeed, I believe that we are entering into a period of research where integration will be more important than introduction. As previously identified within the computer-supported collaborative learning literature, failure to integrate our knowledge of work practice and that of technological development risks increasing an already existing ‘social-technical gap’ (Ackerman, 2000). According to Ackerman, we are increasingly unable to understand and
support what we do socially as opposed to technologically. To address this phenomenon, we must take a more system-oriented view of learner-computer interaction.

To make this point clear, a system-view entails a holistic approach that examines both the theoretical and operational constraints of interaction at cognitive, activity, social organization, and environmental levels. According to Systems Theory, systems that do not possess static linear-like relationships between agents; where components have very little knowledge or understanding of each other; and comprise an open information-space, are classified as complex systems. In this regard, CALL systems are inherently complex. Unfortunately, there is little research in CALL that takes a system-oriented view of learner-computer interaction. Integrating speech technology into language learning systems adds yet another dimension or layer of complexity CALL practitioners must address (Eskenazi, 1999).

Previous research in CALL has recognised the potential benefits of integrating speech technology into language learning environments, ranging from general pedagogy support (Ehsani & Knodt, 1998; Aist, 1999), pronunciation training (Eskenazi, 1999), to speech pathology and applied linguistics research (Esling, 1992). Speech technology in CALL promises to provide an unheralded natural means of communicating at the learner-computer interface, and as such, is an extremely desirable modality for use in interactive systems. Speech in human-human interaction is usable with little training, is the preferred modality of communication, is highly intuitive, and capable of rapidly integrating knowledge acquired from other modalities (such as perceptive, acoustic, haptic). This has led to speech technology at the learner-computer interface as being touted as a means of increasing the authenticity of CASLA activities (Ezkenazi, 1999; Chapelle, 2001). To achieve such a goal, there is need for greater interdisciplinary research. Moreover, we require far greater theoretical understanding of the role and impact of speech technology in CALL than we currently possess. Choice of input modality has previously been shown to affect cognitive information processing (Wong, 2001), however, we still know very little about which modalities compliment particular language learner practices, specifically the role of modality and attention in and even less about the impact of different modal combinations on task performance (Bernsen, 2002).

Chapelle (2001) raises several current and future concerns for CALL practitioners. These concerns can be generally categorised as relating to either development or evaluation practices. However, the discussion in Chapelle provides very little coverage of these concerns in terms of software design and implementation. Furthermore, although previously acknowledging the importance of Human-Computer Interaction (HCI) in CALL (Chapelle, 1997), little is understood about the relationship between learner-computer interaction and effective second language acquisition (Chapelle, 2001:135; Chapelle & Jamieson, 1989). This issue is likely to become increasingly important as practitioners seek to establish more naturalistic interfaces through the integration of advanced language technology in CALL applications (Warschauer & Healey, 1998; Ehsani & Knodt, 1998). Integration of advanced language technology, such as spoken language technology, will undeniably increase the complexity of CALL software development, thus presenting new challenges for CALL researchers and developers. In addition, as CALL practitioners are often ill equipped to effectively integrate advanced technologies into their coursework and teaching practices (Egbert, Paulus & Nakamichi, 2001), investigating appropriate HCI task analysis frameworks should be seen as a critical item on future research agendas.

The present paper reports on a preliminary study investigating the role of multimodal interaction, specifically speech recognition errors, and second language (L2) vocabulary acquisition. The decision to focus on text-to-speech (TTS) and automatic speech recognition (ASR) in the design and deployment of a multimodal learning environment (LIMIT) is based on a growing demand for these technologies in CALL instruction (Ehsansi & Knodt, 1998). This research is motivated by two primary concerns. Firstly, there is a critical lack of such
research in the CALL literature. Second, the existing literature on multimodal interaction and design provides little guidance for the development of speech-enabled CALL systems. Activity Theory is presented as a viable means for evaluating CALL systems due to its ability to incorporate both sociocultural and cognitive theories of learner-computer interaction.

This paper is divided into four sections. The first section provides a general overview of research into multimodal interaction and spoken language technology, and its application to the field of CALL. Subsequently, I provide an introduction to HCI task analysis and discuss its applicability in the design and evaluation of multimodal CALL systems. Note that given the scope of this paper, I do not describe the theoretical and computational aspects of spoken language technology. Interested readers seeking an introduction should see Jurafsky & Martin (2000). For a detailed review of prominent HCI task analysis frameworks, see Diaper & Stanton (2004).

The second section describes an exploratory study into the role of multimodal interaction in second language vocabulary acquisition. In addition to discussing the study’s research design, I present a high-level architectural and functional description of the multimodal-learning tool (LIMIT) developed for the study. Last, this section presents several preliminary observations.

The third section presents an Activity Theory examination of preliminary observations, specifically an analysis of speech recognition errors and their relationship to task performance. Findings from this study greatly contribute to our present understanding of multimodal interaction in CALL system design.

Finally, this paper concludes with a discussion of the implications of this study for future research agendas.

2 Background

Multimodal interaction has received considerable attention over the past decade. Multimodal interaction research is concerned with the integration and synchronization of input and output modalities. The term modality refers to a mode or way of exchanging information between agents (humans or machines) in some environment (Bernsen, 2002). Multimodal systems integrate modalities in a complementary and synergistic manner to overcome the inherent weaknesses of individual modalities. In several instances, multimodal systems have been shown to exhibit greater expressive power, naturalness, flexibility and portability than similar mono or uni-modal systems (Oviatt, 1999a). In this paper I focus on issues surrounding speech recognition as an input modality.

Multimodal interaction design research is principally concerned with two goals, namely achieving natural human-human forms of communication, and increasing the robustness of system interaction (Reeves et al., 2004). Several design categories, or guidelines have been established within the research community to achieve these goals. They are requirements specification, multimodal input and output design, adaptivity, consistency, and feedback (Reeves et al., 2004). Table 1 provides a summary of the purpose of each of these activities. Recently, attention in multimodal interaction research has turned to error prevention, handling and recovery. Research in this area seeks to establish how to minimise the impact of speech errors on task performance, provide users with an appropriate selection of input modalities, and identify when switching between modalities is desirable.

Multimodal interaction in CALL systems is likely to increase substantially with the increased integration of spoken language technology. This is an important area of study as HCI research has found that during task performance, modality integration can often lead to non-optimal user performance and extended task completion times (Grasso, Ebert, & Finin, 1998). This is often caused by the presentation or elicitation of redundant information from one or many input modalities during interaction.
Table 1: Multimodal interaction design guidelines

<table>
<thead>
<tr>
<th>Category</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements specification</td>
<td>Understand how to design multimodal systems for a broad range of users</td>
</tr>
<tr>
<td>Input/Output design</td>
<td>Maximise human cognitive and physical abilities based on human information processing</td>
</tr>
<tr>
<td>Adaptivity</td>
<td>Systems should adapt intelligently to usage and user profiles</td>
</tr>
<tr>
<td>Consistency</td>
<td>System output and performance should be independent input modality type</td>
</tr>
<tr>
<td>Feedback</td>
<td>Users should be made aware of timely contextual information and alternative interaction options</td>
</tr>
</tbody>
</table>

Although each of the categories presented in Table 1 are applicable to CALL, of particular interest to this study are input/output design, feedback and speech error handling.

With respect to input/output design, the context of use must be taken into consideration, as the types of ‘tasks’; users groups; environment; and learning parameters are likely to differ substantially. Input modalities are neither good nor bad, they are more or less apt depending on a particular context. For example, speech recognition systems typically have difficulty recognising spontaneous speech (Robbe-Reiter, Carbonell, & Dauchy, 2000). Human speech production is highly automatic and mostly subconscious. During human-human interaction, interlocutors are primarily concerned with extracting meaning during conversation, and therefore ignore spoken disfluencies, syntactically ill-formed speech and repetition which are poorly handled by today’s speech recognition technology (Oviatt, 1996). Speech modal input is therefore unlikely to be applicable in collaborative language learning environments where spontaneous turn taking occurs. Furthermore, speech recognition is not typically suitable for tasks requiring complex spatial and linguistic references, such as anaphora or deixis (Robbe-Reiter et al., 2000).

Feedback is an essential design consideration in multimodal interaction as attentional resources at the user interface are critical to task performance (McCrickard & Chewar, 2003). The challenge in designing effective multimodal feedback is in reducing distraction (divided attention), whilst providing required information in a timely manner. Although users can support some degree of distraction during task completion, frequent and untimely notifications can impact task prioritisation and increase cognitive load (Robbe-Reiter et al., 2000). McCrickard & Chewar (2003) suggest that effective feedback design must consider frequency of interruption, degree of reactivity required, and importance of knowledge retention.

Although speech recognition accuracy has improved at a rate of 10% over the past ten years (Deng & Huang, 2004), error handling has remained a central concern within multimodal interaction research. Despite improvements in accuracy, error handling is vitally important to design as recognition error rates vary considerably across quiet and noisy environments, recognition engines, and microphone ergonomics. Suhm, Myers and Waibel (1999) found that efficient and graceful error correction was crucial in the design of speech user interfaces. A corollary of their research was the identification that switching between complementary input modalities during speech error correction could improve task completion times and produce more robust system performance. Research has shown that multimodal interaction in speech-enabled environments can lead to improved task performance through superior error handling (Oviatt, 1999b). Oviatt (1999a) outlines several characteristics of multimodal interaction that renders it preferable over unimodal interaction during speech error handling.

1. Users are likely to select an input modality that is least prone to error;
2. A users’ language is often simplified during multimodal interaction, subsequently reducing the likelihood of recognition errors through fewer words and briefer utterances; and

3. Users can employ alternative modes when faced with recognition errors, avoiding continued recognition errors. When faced with complex errors, users prefer reliable over unreliable modes.

Further research is required to examine the veracity of these claims and related research within CALL contexts. Due to the lack of previous comparable research, an exploratory study was determined to be necessary. The purpose of this study is to elicit emerging trends of learner-computer interaction, and to establish an appropriate task analysis framework for evaluating multimodal interaction in CALL.

2.1 HCI Task Analysis

Understanding user characteristics, the context in which users participate in learner-computer interaction, and eliciting both user and organisational requirements for effective system design can be deceptively complex (Farmer, Gruba & Hughes, 2004). This is especially true of CALL contexts in which language learning settings vary widely, and where language learners are often cast as novices, either having very little experience with a particular system or being involved in otherwise novel activities. Simply asking learners what they want, and how they use a system is further complicated by the fact that learners are often incapable of vividly and objectively describing their experiences with the CALL system (Hémard, 1999). CALL applications are used widely for different purposes, in different contexts and under varying conditions, including social constraints (time, place, guidance, learner-autonomy, motivation), which can have a substantial impact upon learner behaviour (Costabile, 2001; Blin, 1999; Hoven, 1999; Levy, 1997). Moreover, learners may possess different educational and cultural backgrounds that may affect their perception of purpose, meaning, and context, which will impact the use of the technology (Gasson, 1999). This makes collecting and analysing whether a CALL system is fit-for-purpose an extremely difficult task. In order to improve CALL system design quality, CALL practitioners and developers need more advanced and holistic techniques for modelling what learners do and what they need. This is the domain of Human-Computer Interaction task analysis. Our first problem is therefore to provide a working definition of task analysis.

The word task is polysemous, conveying several different meanings that are highly dependent upon theoretical background and research domain. Complicating matters, there is no universal definition that is applicable across HCI and CALL disciplines. For instance, a systems theory approach to system design would classify task as an action. A social theory perspective would class tasks as motivated, goal-oriented processes (Soloway, Guzdial & Hay, 1994).

Traditional (cognitive) task analysis in HCI places heavy emphasis on the means/end dimension of context (Kaptelinin, Nardi & Macaulay, 1999). Cognitive task analysis focuses on the performance of a system over time (Diaper, 2004). In this definition, important factors to CALL such as environment, social organization, learner roles and strategies are underrepresented. This becomes increasingly apparent when contrasted with the notion of (communicative) tasks in CALL (Nunan, 1989).

“[A communicative task is] a piece of classroom work which involves learners in comprehending, manipulating, producing or interacting in
the target language while their attention is principally focused on meaning rather than form” (p. 10)

Nunan’s definition of task places greater emphasis on the relationship between meaning and second language acquisition factors such as divided attention and focus-on-form. Although cognitive task analysis does aim to provide a thorough description of individual actions, interactions between actors and other higher levels of activity receive less attention. In measuring performance, cognitive task analysis relies heavily upon discrete empirical evidence. However, effective language learning in computer-mediated contexts cannot always be stated to be a direct corollary of system performance (Hémand, 1999; Hoven, 1999). Long (1997) provides an alternate view of task analysis, where performance is deemed satisfactory as long as it continues to achieve its goals in the application domain. The fact that we consciously undertake a task to achieve some outcome means that pedagogic tasks in CALL must be motivated and goal-oriented (Murphy, 1994), where the overarching goal is for learners to produce language. Furthermore, Murphy states that regardless of its synthetic or authentic nature, a task must act as a catalyst for learning. It may support the idea of mechanical and controlled cognitive activities, or more open-ended and collaborative social activities. This goal-oriented view is representative of the post-cognitive (situated) shift currently occurring within HCI theory.

Generally speaking, the aim of any task analysis framework is to produce a reliable and procedural account of some work aspect for which there is some well-defined practical application. A broader conception of task analysis would seek to establish how people, through the use of socio-technical artefacts, achieve their work goals. This definition applies to both existing and future systems (Kieras, 2004:83). Furthermore, the goal-oriented view of task analysis, rather than the human performance perspective, is appropriate in CALL system design as it corresponds to our previous position that pedagogic tasks must be directed towards some objective outcome (purposeful and motivated action).

Determining the degree to which learners interact with task conditions is fundamental to establishing the task relevancy and theoretical implications of a study (Hulstijn, 1997). An HCI approach to task analysis in CALL must seek to determine the learner’s level of involvement, (possibly competing) goals, actions and learning conditions that are critical to successful system design. The belief presented in this paper is that the dominant cognitive psychological (behavioural) model of task analysis in HCI is limited in its ability to model complex sociotechnical systems, such as collaborative language learning activities. In opposition to the behavioural view of learner-computer interaction, this paper promotes the need for more situated, sociocultural task analysis frameworks in HCI. As such, this paper addresses the paucity of research that has attempted to apply socio-cultural perspectives in HCI to the field of CALL.

2.2 Cognitivist versus Post-Cognitivist Frameworks

Goguen (1996) distinguishes between two dominant theoretical perspectives in HCI, the cognitive (cognitive science and experimental psychology) and social (sociology and anthropology). Traditional HCI frameworks have explored the relationship between human and computer via the discrete modelling of cognitive processes with the intent of embedding cognitive artefacts within the overall system design. However, such approaches fail to recognise the situated nature of activity, and the dynamic emergence of information through mediated agent-artefact interaction (Endsley, 2000; Hutchins, 1995). Situatedness implies that the phenomenon under investigation is situated within some socio-culturally relevant frame of reference. It therefore includes the notion of context, but realises the interdependent relationship between context and tool-mediated activity. It implies situation, that in analysing a phenomenon, we are observing a mere snapshot or immediate instantiation of agent
interaction, and therefore potentially only perceiving a reduced subset of the objective causal mechanisms and the powers that may possibly impact the system (Farmer & Gruba, 2004). Lastly, the term situatedness implies an ecological perspective, that the phenomenon exists within a specific environment, and that the nature of this environment affords particular relationships and mechanisms to which actors may respond (Gibson, 1966; 1977). The view that participation within an activity is not isolated, yet inexorably bound to external relationships with other actors, artefacts and socio-cultural conditions poses serious questions for how software engineers, human factors experts, and CALL task designers proceed collaboratively with requirements engineering, system design, and evaluation.

Contrary to traditional cognitivist views of HCI, situated theories are driven by sociological considerations – how work is shared, organised and completed within communities of practice (Preece et al., 1994). Supporting existing work practices within social contexts requires more than understanding a single individual’s perception of a task (Ackerman, 2000). Rather, it is necessary to understand a learner’s actions within the context of situated activity, recognising the importance of modelling existing environmental artefacts and conditions, subject (individual or group) collaboration and cooperation, and motivation (Hutchins, 1995; Suchman, 1987; Engeström, 1987; Nardi, 1996). Growing trends in HCI research to establish better contextuality of interaction; greater understanding of agents acting in situ, and how agent-artefact interaction (tool-mediation) constructs knowledge (Bannon, 1991) are representative of this post-cognitivist paradigm shift (Kaptelinin & Nardi, 2003).

2.2.1 A Situated Perspective: Learner-Centred Design and Activity Theory

Where modelling activities are constrained by socio-cultural conditions, such as language learning (Hoven, 1999), the task analysis framework must be able to support the dynamic and emergent nature of social interaction and resist the temptation to model computer-mediated activity as the static, simplistic collection of cognitive actions (Albers, 1998). Rather than focusing on performance, system design and evaluation in CALL should be more learner-centred, concentrating on supporting learner practices, such as the acquisition of expertise and learner engagement in novel learning situations.

This learner-centred design (LCD) approach is more concerned with eliciting and elucidating important factors in CALL such as learner roles and responsibilities, SLA task structure, tool or artefact mediation, shared knowledge structures and cultural conventions. Most importantly, LCD considers how learners acquire meaning from their interaction with the environment. Furthermore, LCD approaches should not be overly concerned with reducing the complexity and richness of the interaction for the purpose of improving performance, as conflicts and contradictions in learner perceptions and beliefs (breakdowns) represent unique learning opportunities for the development of new meaning and strategies (Nardi, 1996; Leont’ev, 1978; Engeström, 1999).

LCD in CALL has been linked to the importance of self-managed individual learner strategies (Hoven, 1999), whereas HCI perspectives on LCD focus on identifying those traits that distinguish a learner from a typical user (Quintana, Carra, Krjcik, & Soloway, 2001; Soloway et al., 1994). LCD seeks to establish the ecological validity of contextual use that constrains learning activities. Indeed, as noted by Chapelle (1994), the context in which learning occurs and the nature of collaboration between learner agents (Human-Human Interaction) and artefacts of the environment (Human-System Interaction), directly impacts the production of language in CALL environments. According to Hoven (1999), language production in CALL implies negotiation, mediation, and interaction. Task analysis in LCD requires an integrative approach that draws upon the strengths of cognitive and sociocultural accounts of language learning (Kumaravadivelu, 1983). LCD in CALL requires a more
appropriate unit of analysis than *cognitive tasks* for investigating learner-computer interaction. In sum, LCD in CALL must ensure that system design and evaluation is:

4. Context sensitive;

5. Focused on situated activities framed within a cultural-historic view;

6. Goal-Oriented; and

7. Considers the dynamic flow of information between actors and artefacts during computer-mediated activity.

At present, LCD is a metaphor for describing theoretically sound principals for investigating learner-computer interaction in CALL. Already, describing the types of factors involved in LCD proffers some initial benefits, yet for LCD to become an integral and effective component of CALL system design, a more detailed methodological framework is required. Activity Theory is promoted in this paper as one such viable framework.

Activity Theory has a long and rich history. Steeped in socialist and transcendental theory, Activity Theory was derived from the socialist thinking of Kant, Marx, Hegel and Engels (Kuuti, 1996). Activity Theory has received considerable contributions from Vygotsky (1986), Leont’ev (1978, 1981) and Engeström (1987, 1999). Rather than a scientific theory of interaction, Activity Theory is a descriptive conceptual framework. As such, its purpose in HCI is not to empirically evaluate computer-mediated activities, but to describe the different forms of human praxis and developmental processes involved in human-computer interaction. As for practical applications in CALL system design, Activity Theory can be used to inform future system designs, establish learner roles, investigate collaboration and cooperation strategies (including negotiation), and clarify situational and environmental constraints. Although Activity Theory has only recently been applied to the field of CALL (Blin, 2001), it has been widely applied to Information Systems, Computer-Supported Collaborative Work and HCI research (Turner, Turner & Horton, 1999).

As a situated task analysis framework in HCI, Activity Theory provides a much broader theoretical framework than competing cognitivist frameworks for examining computer-mediated collaborative activities (Stanton, 2004). As the name suggests, the unit of analysis in Activity Theory is the activity itself. Activity Theory is highly relevant to task analysis for LCD in CALL due to its relationship with Vygotsky’s *Zone of Proximal Development* (Vygotsky, 1986), indicating the possible range of actions a learner may perform whilst collaborating on some computer-mediate learning activity. In Activity Theory, individual and social processes are interwoven in the generation of contextualised meaning. Individual knowledge and meaning cannot be separated from the situatedness of the activity from which they are generated.

Activity Theory is an object-oriented, as opposed to goal-oriented, task analysis framework. The importance of goals is present, however they are subordinate to motivation, which is always oriented towards some objective purpose or outcome (object) (Whittaker et al., 2000). Accordingly, interaction in learner-centred environments is considered purposeful. Hence, system designs based on activity-centric analyses are more likely to address the language learner’s motivation, than focus on discrete tasks (actions) alone (Whittaker et al., 2000). Activity Theory implicitly handles the frequent changes in context experienced by language learners as the activity defines the context, and the context defines the activity. Therefore, the act of language learning in one context should not be considered congruous with the activity of learning in another.

Activities are always real-life situations, and thus impossible to understand without some contextual reference (Kuuti, 1996). Activities are not static models of interaction; they are dynamic and continuously change according to fluctuating conditions and transforming
goals. Activities are mediated, individuals carrying out their tasks through the use of available artefacts in the environment. When an artefact is used to mediate the subject’s purposeful interaction in the activity, it is said to become a tool. A tool can be any material or mental process used to transform or convey information. Tool mediation is thus one of the central tenets of Activity Theory. As artefacts are themselves produced through some activity, they contain cultural and historic references. Artefacts, as tools in an activity are therefore capable of imparting external cultural and historic beliefs and conventions upon the current activity.

Activity Theory posits a hierarchical structure of activity, consisting of three primary components: activity, object-oriented actions, and operations (Leont’ev, 1978; 1981). As Figure 1 indicates, the first layer within Activity Theory is the activity itself, which is motivated by the subject (individual or group). An activity is comprised of mediated object-oriented (motivated) actions. Associated with each action is one or more parent and sub-goals. The fact that each action may be associated with one or more goals infers that actions are polymotivated, meaning that the same action (task) may be applied to accomplish different goals. This is important as it provides a way of supporting alternative paths or strategies within an activity for obtaining the objectified outcome. Additionally, it helps to explain why we cannot focus purely on the action itself, as its purpose or role is ambiguous. Lastly, operations are seen as unconscious actions, those behaviours that have been habituated through experience.

Figure 1: Activity Theory framework

As suggested, Activity Theory posits that there is constant interaction between these layers within the activity. It must be noted that a more recent formulation of Activity Theory proposed by Engeström (1999) provides additional support for analysing collaborative work practices by also considering community, division of labour, and praxis within the model.

Using Activity Theory we can also start to investigate how actions are related to goals and how they possibly impact motivational aspects of activity involvement. Specific to CALL, this process may be helpful in identifying how changing conditions in language learning tasks increase or reduce cognitive load during learner-computer interaction. Although all levels of activity can move up and down in the structure hierarchy, the most likely transitions are to occur between operations and actions, only severe conflicts or breakdown resulting in changes to the motivational nature of the activity. This movement also helps to analyse the severity of the identified problem.

3 LIMIT: A multimodal learning tool

In order to examine the relationship between multimodal interaction with speech technology and second language vocabulary acquisition, it was necessary to develop a new multimodal learning environment. LIMIT (Learning Interactively through Multimodal
Instruction) is an extensible and portable Java-based application that can run in either a local or distributed environment. LIMIT incorporates both TTS and ASR functionality, as well as advanced event-based multimodal activity logging. LIMIT was developed with several important quality attributes in mind, notably ease-of-use, scalability, portability and extensibility. Table 2 provides a description of these quality attributes.

**Table 2: Quality attributes of the LIMIT architecture**

<table>
<thead>
<tr>
<th>Quality attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease-of-Use</td>
<td>Built with rapid lesson development in mind, the user interface, speech technology (languages, grammars, etc.) and error handling is derived from the XML lesson file. Teachers are easily able to construct new multimodal learning environments by simply providing new lesson content.</td>
</tr>
<tr>
<td>Scalability</td>
<td>As CALL environments are becoming increasingly network-centric, LIMIT was developed to adapt to both communicative and integrative CALL requirements. Therefore, LIMIT may be used in the classroom with one or many users.</td>
</tr>
<tr>
<td>Portability</td>
<td>Built with Java and XML, LIMIT can easily be deployed on Windows, Macintosh and POSIX operating systems.</td>
</tr>
<tr>
<td>Extensibility</td>
<td>As requirements frequently change, LIMIT was developed to be highly modular. New functionality can be rapidly integrated into the framework.</td>
</tr>
</tbody>
</table>

Figure 2 provides a high-level architectural view of LIMIT. Lessons are written in XML. A separate application (not described here) may be used to assist instructors to construct lessons, which are then transformed into the required XML format. Apart from the general benefits of developing with XML in CALL (Mishan & Strunz, 2003), XML provides a low-bandwidth means of providing learning and configuration information in network-centric environments. Having created the lesson file, it is then read and parsed within LIMIT to construct the question and answer set. The structure and content of this set is then used to derive and dynamically construct an appropriate Graphic User Interface (GUI). In addition, based on the content of the lesson, LIMIT establishes an appropriate connection with an available speech recognition and text-to-speech server. LIMIT supports both open-continuous (open microphone, sentences) and closed-discrete (click-to-speak, isolated words) recognition. When more than one ASR and TTS server is available, learners are presented with the opportunity to select a suitable server based on their preference for language and accent models. A general benefit of this approach is that the application can intelligently adapt to the type of lesson provided by the instructor. This subsequently reduces the complexity of constructing speech-enabled language learning activities. Keymodal, graphic manipulation and speech events are all recorded and logged to a file for subsequent statistical analysis. All events are time-stamped with start and end times, which provides a basic means of conducting time-series analysis of the data.
As shown in Figure 2, LIMIT can be used locally or over a network with ASR and TTS servers. Table 3 presents ASR and TTS engines proven to work with the LIMIT architecture. In general, LIMIT can communicate with any speech engine that provides a Java interface or is SAPI 4/5 compliant.

<table>
<thead>
<tr>
<th>Speech Recognition Engines (license)</th>
<th>Text-to-Speech Engines (license)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuance Version 8.0 (commercial/developer)</td>
<td>Vocalizer Version 1, 2 (commercial/developer)</td>
</tr>
<tr>
<td>CMU Sphinx (open source)</td>
<td>Festival (open source)</td>
</tr>
<tr>
<td>Dragon Naturally Speaking (commercial)</td>
<td>MBROLA (open source)</td>
</tr>
<tr>
<td>L&amp;H Microsoft Agent Voices (commercial/developer)</td>
<td></td>
</tr>
</tbody>
</table>

Finally, LIMIT provides a configurable level of feedback to learners ranging from iconic representation of activity (microphone is on/off) to a detailed view of keystroke, mouse, TTS, and ASR events. Tailored feedback for incorrect learner responses can be provided by instructors in the lesson file, or dynamically generated by the system.

4 An Exploratory Study
Due to a critical lack of research in CALL, a small qualitative, exploratory study was conducted to investigate the role of multimodal interaction in second language vocabulary acquisition. This study was deemed necessary to establish an initial research agenda and to
verify and validate methodological research tools (Reis & Judd, 2000). The LIMIT-based study was undertaken at a leading Australia university. The aim of the study was to explore the impact of input modality selection upon immediate and delayed recall and recognition of novel second language (L2) lexicon.

4.1 Procedure

Instructional material for the study was based upon the CAVOCA framework presented in Groot (2000). Subjects were asked to participate in an eight-week French language learning program. This program consisted of three LIMIT-based sessions over two weeks, followed by a delayed recall and recognition test six weeks later. Participants \((p=4)\) were randomly assigned evenly to two groups. Groups differed in their access to speech recognition (both had access to TTS services). Participants with access to speech recognition were given complete autonomy in their choice of input modality, choosing if and when they decided to interact via speech recognition. In order to reduce the likelihood and impact of speech recognition errors, click-to-speech functionality was used to enforce a closed-discrete speech recognition interface. Once participants had ‘clicked to speak’, recognition occurred via voice onset and offset timing. In the event that the participant lapsed back into a more natural style of speaking, extensive grammars were provided to extract salient words within an utterance and to minimise the impact of disfluent or repetitive discourse.

Pre-study questionnaires were used in the participant selection process to reduce the possibility of including primed individuals (through previous exposure to French) in the study. Targets L2 words (items) were selected based on the following criteria:

- Frequency of use;
- Level of complexity;
- Likelihood of priming; and
- Nouns

These criteria correspond to (a) the importance of determining item complexity when selecting L2 items based on vocabulary frequency tables (Chapelle, 2001:136), (b) the impact of semantic priming upon recall ability (McNamara, 1992), and (c) evidence from child and cross-cultural SLA research that indicates greater recall ability of nouns over adjectives (Gasser & Smith, 1998).

4.2 Tasks

Each participant session with LIMIT was structured around the CAVOCA framework. In each session, participants were asked to respond to a series of L1 questions pertaining to each L2 word. These questions were structured upon the principles of task-based learning (Groot, 2000; Skehan, 1998), which were broken down to three distinct stages: inference, elaboration, and consolidation. During inference, an L2 word is introduced to the learner, followed by a series of multiple-choice questions. During elaboration, learners are asked to reason about the logical validity of a statement using the same L2 word. In consolidation, learners are presented with a text in English and in French and asked to associate the related L1 and L2 words. This process is then repeated for all words in the exercise. Despite the relative simplicity of these tasks, the motivation behind the design was to cater for novice L2 learners with little or no exposure to French. Responses to questions were verified against provided answers, logged, and immediate feedback provided to the learner.

To ensure the collection of rich accounts of learner-computer interaction, the study employed a number of traditional ethnographic research tools (Punch, 1998). Pre-study questionnaires and semi-structured interviews were used to elicit background technical and sociocultural information about each participant. During learner-computer interaction, the
researcher directly observed participants, and in addition to the quantitative logging of computer-based events, qualitative data was manually recorded through note taking and audio recording. This additional data was subsequently used to validate the results of the multimodal logger (were recognition errors correctly captured?) and to analyse both verbal and non-verbal behaviour. Verbal (Think Aloud) Reports were used during the study to provide an enhanced level of acuity into the cognitive processes of the user during task completion (Ericsson & Simon, 1993).

4.3 Assessment

Upon completion of each session, learners were asked to participate in an immediate post-session recall test. This test consisted of L1 equivalents of the L2 target items. Given the L1 equivalent, learners were asked to provide the L2 equivalent. Results were recorded. A small semi-structured interview then took place to acquire additional feedback from the participants. Finally, after a six-week interval, participants were asked to participate in a delayed recall and recognition test. The structure of the recall test was consistent with that of the immediate post-session test. In the recognition test, participants were provided with a series of multiple-choice L1 questions in which the correct L2 word was to be selected from a list of alternatives. Responses were recorded and compared with previous results from the initial three sessions.

5 Preliminary Observations

In this section I report on L2 vocabulary retention results from the study as well as several of initial observations concerning multimodal interaction design and the impact of speech recognition errors on learner performance. Firstly, I report on observations concerning multimodal interaction design.

5.1 Multimodal Interaction Design

Participants in the speech recognition group demonstrated several characteristics previously recognised in the research literature (Oviatt, 1999a; Suhm et al., 1999). Speech recognition proved to be extremely robust with only one identified false recognition across the six speech sessions. This can be accounted for by three factors. First, recognition was closed-discrete, with a minimal main dictionary and supported by grammars. Therefore, isolated word recognition was more robust. Second, sessions took place in a quiet laboratory setting, relatively devoid of ambient noise. Third, false recognitions were likely offset by the learning parameters of the confidence limits that captured ambiguous input and prompted for clarification before returning a final result. This feature was found to be extremely useful as clarifications were significantly reduced as learning parameters adjusted to the participant.

Although participants were not instructed to use speech, it was their preferred modality, switching to keymodal or graphical manipulation input modalities for only in instances of repetitive speech recognition clarifications. Time for task completion was significantly different between the two groups; speech-modal participants taking on average three times as long to complete the session exercise. Task time also increased with respect to error handling events. Additionally, with each repetition, speech recognition interaction time increased, participants displaying more rehearsed and articulate speech. Text-to-Speech (TTS) was also used substantially more during error handling. This can be seen as an instance of self-instruction, each participant using TTS to train their speech to that of the computer’s. In general, participants from both groups found the TTS services extremely useful and engaging.

Input modality integration was greatest during error handling. Non-speech participants did not significantly vary in their use of keymodal and graphic manipulation modalities. Speech-modal participants demonstrated a high degree of modality switching,
indicating the multimodal interaction is a successful and robust means of overcoming input modality failures. For both groups, keymodal and graphic manipulation were the efficient input modalities. However, speech-modal participants displayed continued preference for speech modal interaction.

A particular behaviour with speech-modal participants that was unexpected was the use of speech recognition confidence level feedback by learners to evaluate their performance. Both participants used the confidence scores to establish whether their pronunciation was improving with each stage, as well as evaluating the performance of the recognition engine. This indicates that feedback indicating speech recognition accuracy can provide an additional form of useful feedback during multimodal interaction.

Direct observation, analysis of the verbal reports and transcription of audio recordings revealed that no user interface design difficulties were encountered. In addition, participants felt the system feedback was both sufficient and appropriate.

5.2 Second Language Vocabulary Acquisition

The full extent of the importance of this study was not determined until after the delayed recall and recognition testing took place. Here, it was found speech-modal learners possessed a more acute recall ability than non-speech modal learners.

5.2.1 Immediate Post-Session Testing

Non-speech modal participants achieved 100% recall of the presented novel L2 vocabulary. Although it was not a requirement, orthographic representations were fairly accurate. Additionally, participants were able to recall the pronunciation of each word with reasonable accuracy. These results can be accounted for considering this groups primary use of keymodal input (text) in conjunction with TTS interaction.

The speech-modal group displayed 40% recall ability of the presented novel L2 vocabulary. However, for all words recalled (different for both participants), orthography and pronunciation were accurate. Post-session interviews and verbal reports indicated that recall for these target items was related to words that required the greatest amount of speech error handling during the session. This fact was confirmed via subsequent analysis of the multimodal logger data. Analysis of the logs also indicated increased TTS and keymodal events for the relevant target items.

5.2.2 Delayed recall and recognition testing

Both non-speech modal participants experienced a 60% reduction in recall ability of the previous presented L2 vocabulary. By contrast, speech-modal participants experienced a 0% reduction in recall ability. For recognition, both speech-modal participants experienced a 100% improvement over delayed L2 vocabulary recall ability. On the other hand, both non-speech modal participants experienced near 100% perfect recognition all the same L2 vocabulary items.

In conjunction with immediate post-session recall tests and the various forms of user accounts of interaction, it was possible to establish an emerging causal relationship between input modality and L2 vocabulary retention. Furthermore, applying both GOMS and HTA representations to the analysis of the recorded qualitative and quantitative data could not explain this relationship.

6 Discussion: An Activity Theoretic Analysis

As application of traditional cognitive HCI task analysis methodologies was unable to establish a theoretical foundation for the emerging causal relation between multimodal interaction and L2 vocabulary acquisition, attention was turned towards a more sociocultural
approach. Activity Theory was chosen for its ability to incorporate both cognitive and sociocultural representations of interaction (Kaptelinin, 1996). Given the findings of the initial post-session recall tests, the activity theory analysis undertaken was limited to speech recognition errors and multimodal interaction.

According to Activity Theory and the principles of learner-centred design, motivation and consciously maintained goals are critical for task analysis in learner-computer interaction. Referring to Figure 1 we see that motivation is central to continued participation in some activity. As the inherent nature and structure of the activity did not change throughout the course of the study, subject motivation can be said to have remained stable. During interaction, speech-modal participants maintained a series of conscious goals that were motivated by the nature of the activity and which subsequently directed their actions. However, when speech recognition errors occurred, these introduced conflicts and contradictions into the activity. These conflicts according to Activity Theory result in the raising of previous unconscious operations (habituated actions) to consciously maintained goals. This has the effect of modifying the individual’s situational awareness of the activity (Suchman, 1987; Kaptelinin, 1996). In the case of speech recognition errors, this resulted in additional attentional resources being directed towards the task of error handling and correction. As human-human speech communication is largely automatic and unconscious (through habituation), it can be assumed these conflicts led to increased cognitive processing of the cued-stimulus at the user interface (namely, error handling feedback).

Activity Theory suggests that conflicts and contradictions lead to breakdowns. Unlike cognitive accounts that posit how errors constitute a fragmentation of the learning process (Skehan, 1998), an activity theoretic perspective perceives these error states as learning opportunities. Therefore, these breakdowns result in greater attention to task, increased focus-on-form (unconscious operations $\rightarrow$ conscious goals $\rightarrow$ increased cognitive processing), and increased opportunity for cognitive apprentice (Vygotsky, 1986). With respect to cognitive apprenticeship or expertise, interlocutors (in this case, learner and computer) communicate with the mutual goal of understanding one another. Both partners negotiate the meaning of their message by making interactive modifications (negotiations) aimed at achieving mutual comprehension. These negotiations represent learning opportunities that ultimately lead to greater expertise and mastery of the second language.

As the L2 target items recalled by both speech-modal participants received the greatest amount of interaction, due to increased difficulties in error recovery, it can also be assumed that a greater amount of cognitive processing of the L2 target items occurred. This view is consistent with the findings that on these particular target items, speech recognition time for both participants increased with each error recovery attempt. This was not the case when speech recognition errors did not occur. In these instances, we assume that habituated operations remained largely unconscious and automatic, requiring fewer attentional resources and less cognitive processing of cued-stimulus. Assuming also a correlation between goal-orientation and focus-on-form, we can account for the poor immediate post-session recall results by suggesting a lack of sufficient awareness and attention to the lexical and phonological structure of the L2 target item. In other words, focus-on-form was not a consciously maintained learner goal.

In sum, activity theory provides an accessible account of dynamic, non-linear information flow during multimodal interaction in CALL. Moreover, it presents the ability to describe L2 acquisition in terms of motivation, goal-orientation and conflict resolution, hence remaining commensurate with learner-centred design principles and constructivist language learning theories.
6.1 An SLA and Information Processing Perspective

Information processing models of language acquisition further support the previous goal-oriented analysis provided by Activity Theory, suggesting a relationship between motivation, conscious goal awareness, attention and language acquisition. Information processing frameworks such as Levels-of-Processing (Craik & Lockhart, 1972) and Connectionism (Collins & Loftus; 1975) have been used to demonstrate a relationship between long-term retention of declarative-semantic memory, depth of information processing, and attentional resources. In the case of SLA, levels-of-processing suggests that the deeper the processing of novel cued-stimulus, the greater the likelihood of L2 lexicon retention. This is especially true of echoic and visuals cues (speech and images) (Craik & Lockhart, 1972).

Three hypotheses emerged from the previous Activity Theory analysis of speech recognition errors and L2 vocabulary retention. These were,

1. Speech recognition errors do not immediately constitute learning inhibitors;
2. Multimodal interaction can have a positive impact on L2 vocabulary acquisition; and
3. Allocation of attentional resources and situation awareness during multimodal interaction are essential for long-term L2 acquisition.

Both of these hypotheses find support within the information processing literature. Firstly, multimodal interactive environments typically lead to a greater number of event stimuli being provided during semantic encoding (Oviatt et al., 2000). The transference of multiple stimuli per target item will consequently result in increased preservation of neural pathways (Collins & Loftus, 1975). As increasing information processing requirements correlates to a greater demand on attentional resources (Long, 1988; Skehan, 1998), there is a likely relationship between multimodal interaction, attention and language acquisition.

Analysis of attentional resources during multimodal interaction can also be used to describe the apparent disparity between immediate and delayed recall ability amongst participants using speech recognition. Speech recognition introduces a novel input modality for most users, which requires additional cognitive processing. Furthermore, selection and coordination of input modalities typically increases task complexity (Grasso et al., 1998). It is expected that during normal task execution that multimodal interaction will result in a higher degree of divided attention and reduced focus on form, therefore significantly reducing information retention (Craik & Govoni, 1996; Baddeley, Lewis, Eldrige, & Thomson, 1984). However, during error recovery or input disambiguation, participants were required to attend to fewer tasks. Given the nature of interaction, they were also prone to be more attentive and aware of a single task. As recovery and disambiguation requires further elaboration of the target item, the ‘quality’ of attention is likely to result in higher rates of attention (Lockhart & Craik, 1990). Also, Wixted and Ebbesen (1991) found that although forgetting occurs at a constant (logarithmic) rate, initial increases in attention and information processing lead to an increase in retention. Therefore, the disparity between initial and delayed recall ability for speech recognition participants can be accounted for by the increased level-of-processing, attention, awareness and elaboration of the target item during speech recognition error recovery and disambiguation. The inverse can be said for the non-speech group. This is in direct opposition with previous finding by Clark & Clark (1997) suggesting that attending to errors during learning diminishes the individuals capacity to accurately retain and subsequently recall novel information.
Conclusion

The research reported in this paper contributes to the field of CALL by demonstrating that selection of input modalities in language learning environments can affect various aspects of SLA, including delayed recall and recognition of novel L2 words. Results from the study are also consistent with previous research demonstrating the potential benefits of multimodal interaction accuracy, efficiency, flexibility and ease of error handling (Oviatt, 1996). Moreover, this paper has described how an activity theory analysis of multimodal speech recognition errors in CALL contexts can provide additional insights into learner-computer interaction. However, much research remains to be undertaken.

Previous discussions on CALL software development models and frameworks (Hubbard, 1996; Levy, 1999) fail to inform CALL practitioners as to appropriate practices and procedures for evaluating learner-computer interaction. Hence, I have argued in this paper that increased task analysis research in CALL is required. Additionally, I have argued that task analysis frameworks in CALL must be capable of capturing and evaluating both cognitive and sociocultural aspects of learner-computer interaction. The preliminary findings from an examination of multimodal interaction in second language acquisition suggest that traditional cognitive task analysis methodologies are poorly suited to evaluating computer-assisted language learning environments. In contrast, it has been shown that Activity Theory provides a flexible and accessible framework for responding to the dynamic and complex characteristics of CALL system design and evaluation.

In concluding, it is appropriate to mention limitations of the present study and the application of Activity Theory. Firstly, the small sample size and nature of the study reported in this paper precludes any longitudinal analysis of the emerging factors related to multimodal design and second language acquisition. Although eighty percent of human-factors related information can usually be obtained using 8-12 participants (Nielsen, 1993), the sample population in this study was both quantitatively and qualitatively too small for ensuring the generalisability and transferability of findings. However, with preliminary findings suggesting several emerging trends for future research, a larger focused qualitative study is currently being conducted to re-examine the results reported herein.

Second, although it has been shown that Activity Theory is capable of evaluating learner-computer interaction in CALL contexts, it has yet to be integrated into a more holistic framework that supports rigorous software engineering efforts. Frameworks that build upon the principles of Activity Theory, such as CASE (Farmer & Hughes, in press), promise to greatly improve the overall quality of CALL system design.

Finally, the present study examined individual and subjective accounts of learner-computer interaction within laboratory settings. However, speech recognition performance, specifically error rates, differs dramatically between quiet and noisy environments and across recognition systems (Deng & Huang, 2004). The effective design of multimodal CALL environments is therefore greatly dependent upon the type of activity, especially individual or collaborative styles of learner-computer interaction. Therefore, we require additional research regarding the nature of multimodal collaborative CALL environments.

Given the wide range of issues yet to be treated by the research community, continued research into the role of multimodal interaction in CALL is required before guidelines for effective multimodal CALL system design can be produced. Open areas of research relevant to CALL include, but are not limited to, user profiling, modality selection/combination theories, multimodal interaction and task performance, and adaptive speech environments.

References


ERROR: undefined
OFFENDING COMMAND:

STACK:
Author/s: 
Farmer, Mr Rod

Title: 
Multimodal speech recognition errors and second language acquisition: An activity theoretic account

Date: 
2005

Citation: 

Publication Status: 
Published

Persistent Link: 
http://hdl.handle.net/11343/33818

File Description: 
Multimodal speech recognition errors and second language acquisition: An activity theoretic account