Umeå University
Department of Computing Science
SE-901 87 Umeå, Sweden

UMINF-14.07
February 2014

Does TTS-based Pedestrian Navigation Work? [1]

by

Michael Minock, Johan Mollevik and Mattias Åsander

[1] The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270019 (SPACEBOOK project www.spacebook-project.eu) as well as a grant through the Kempe foundation (www.kempe.com).

ABSTRACT

We seek to test the hypothesis that *text-to-speech(TTS) navigation systems can adequately guide pedestrians to unknown destinations in an unfamiliar city*. Such systems bypass screen-based, multi-modal techniques and simply speak route following instructions incrementally into the pedestrian's ear piece. Due to errors in GPS positioning, uncertainty of user heading, poor map quality and potential communication and processing latencies, this becomes a surprisingly challenging task. In our study, subjects are led on an unknown tour on the grounds of Umeå University. We evaluated both a human wizard controller as well as a simple decision-tree based controller and compared them to an ideal subject that knows the route. Results give support to our hypothesis that TTS-based navigation systems can adequately guide pedestrians. That said, our experiences point toward immediate and future improvements to make such systems more effective and agreeable.

All the software and data behind this work will be open sourced to encourage confirmation, replication and, ultimately, improvement upon our results. This will soon be available for public download at `http://janus-system.eu`.

1 Introduction

We are witnessing an upsurge in mobile applications that help pedestrians navigate cities. Commonly such applications are map-based, incorporating user interface techniques adapted for small touch screens. While such approaches are workable, they can be unsafe and awkward in busy urban environments; pedestrians should endeavor to keep their eyes and hands free. Instead of adopting exotic solutions, such as haptic interfaces [14] or displays built into eye glasses, we assume a more basic set-up: *the user only hears route following instructions through their ear piece while their mobile phone, in their pocket or purse, tracks position.*

To our knowledge this 'ears only' set-up was first introduced in the EARS project [1], though that project did not address navigation and the server was a laptop in the user's backpack. In the work here, the focus is on navigation where the phone application communicates with back-end components through the 3G network.

While much has been learned through natural language generation applied, one shot, to generate instructions for entire routes [4], we focus on the case where shorter instructions are presented incrementally. Also, while many incremental systems work over simulated environments [10], including those that generate noise to simulate GPS errors [6], we focus on fielded systems that must handle real GPS errors and network latencies – data that does not easily conform to standard distribution models. Furthermore while strong evidence supports the adequacy of open data [3] for the vehicular case [5], in the pedestrian case the system must be able to describe salient landmarks [15] at a finer scale, and infer or represent traversability across parks, squares, etc.

Pedestrian navigation systems based on 'ears only' set-ups have only recently been developed [9, 2, 12] and evaluated [7, 2]. This report performs the first study comparing wizard and automatic guidance against one another and against an ideal walker who knows the route. Such a study should give direct evidence to the question that is the title of this report. While it is possible that similar efforts are currently underway (or soon will be) in industrial labs, the effort here is emphatically 'open'. All the software and data behind this work will be open sourced to encourage confirmation, replication and, ultimately, improvement upon our results.

Organization of this Report: Section 2 describes our experimental design. Section 3 describes the simple reactive controller we used in the evaluation. Section 4 presents our results. Section 5 discusses some of what we learned and our plans for future research, development and evaluation. Section 6 concludes.

2 Design of Experiment

Our experiment seeks to test the hypothesis that *TTS-based navigation systems can adequately guide pedestrians to unknown destinations in an unfamiliar city.* While this assumes that the subject is in a unknown environment, in reality, this is generally not the case – subjects, often students, must be recruited in the test city. Thus we observe the following constraint: *Using street and other place names is prohibited and all landmark references are restricted to physical descriptions in the subject's immediate visual environment.*

A single trial in our experiments guides a subject, recruited via a flier, on a 10 minute walk, followed by a short exit interview. Subjects are greeted and asked if they have heard anything about the project or experiment. After we establish that they are fresh subjects, the subject dons the device and is guided outdoors to a starting position. Once the GPS antenna connects, the subject starts walking down a path. At this point the timer starts and the controller attempts to guide the subject to a sequence of goals for 10 minutes. Figure 2 shows the route used in our study with nine goals marked by red dots. The destinations are staged so that the first should be relatively easy to reach with the following destinations requiring more sophistication. When a subject reaches a goal, they are asked a yes/no question of whether they see the goal. The subject then continues on the tour to the next goal. After 10 minutes the subject is told to return for debriefing and to receive a 'free' lunch coupon.

3 Controllers

In this report we conducted evaluations under two types of controllers: 1.) A human wizard controller; 2.) a simple reactive controller. Both controllers are deployed in the infrastructure described in [12, 13].

3.1 The Wizard Controller

The wizard controller is simply a human expert, aware of the tour, sending text messages to the subject in real time to guide them onward. In these trials the subject wears the device as a necklace and the wizard can view a real-time stream of images taken from the phone camera. The wizard is also helped by hearing a real time audio stream and seeing GPS positions updates in real time on a map.

3.2 The ASAP Reactive Controller

Our reactive controller, which we call the ASAP (for *as-simple-as-possible*) controller, is the simplest controller we could imagine that might succeed. This controller does not engage in dialogue; the controller talks to the pedestrian, but the pedestrian does not talk back. The controller queries the pedestrian state each second. This state includes facts such as which elementary edge and segment the pedestrian is on, if any, and which branching point they are approaching, if any[1]. This includes facts about heading, heading correction, etc. Error measures on headings, position, speed are also calculated based on the time series of prior measurements.

Given this detailed information, every second the controller determines a true/false value for each of the following seven propositions: p_1:`receiving-tts` (is the subject currently being spoken to?), p_2:`gps-adequate` (are recent GPS reports accurate enough to determine current position?), p_3:`at-goal` (is subject at current goal?), p_4:`on-path` (is the subject on a path-segment which can reach the goal?), p_5:`at-branching-point` (is the subject at or approaching

[1] Our terminology is largely based on that of [16].

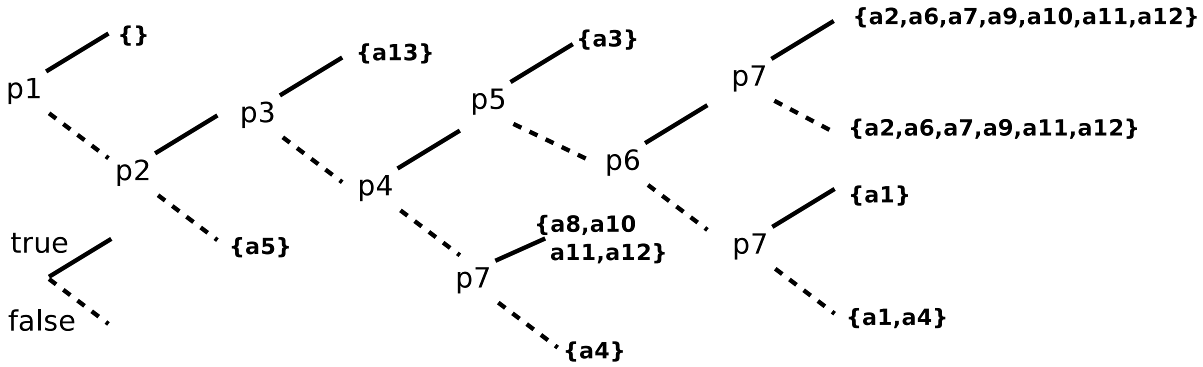

Figure 1: The decision tree determining applicable speech acts

a branching point?), p_6:`aligned-with-edge` (is the subject's heading aligned with the current elementary edge of the path-segment that they are on?) and p_7:`heading-accurate` (is the time series of recent GPS stable enough so that heading may be relied on?).

Based on the truth values of the seven propositions, the controller decides the next utterance to produce, if any. We model this in 13 speech acts encoded in the meaning representation language (MRL) described in [17]. The speech acts are divided into the categories of *Immediate instruction*, *non-Immediate instruction* and *Inform* acts. The Immediate instructions are a_1:`instr-heading-correction` ("orient slightly to your left"), a_2:`inst-continue` ("continue down the gravel path between the two buildings"), a_3`inst-immediate-turn` ("turn right!"), a_4:`instr-move-for-heading` ("walk in a straight line so that we can calculate your heading"), a_5:`instr-wait-for-GPS` ("stand by for better GPS measurements"). The non-immediate instructions are a_6:`instr-continue-on-span` ("keep walking on this path for 200 meters"), a_7:`instr-future-turn` ("in 60 meters, veer left toward the fountain in the middle of the square"). The inform speech acts are a_8:`inf-euc-distance` ("you are 200 meters, as the bird flies, to the goal"), a_9:`inf-path-distance` ("following the route, you are 300 meters from the goal"), a_{10}:`inf-heading-to-goal` ("the goal is at your 10 o'clock 200 meters away"), a_{11}:`inf-recent-progress` ("you seem to be walking away from the goal"), a_{12}:`inf-landmarks` ("the large two story white building to your left is Universum"), a_{13}:`inf-goal-reached` ("you have reached the goal, do you see a bank machine?").

The decision tree in figure 1 indicates which speech acts are available for each state of the system. More than one speech act may be triggered, thus the controller must pick which one to generate and present, if any. We enforce a 10 second period between presentations, with *Immediate instruction* speech acts being able to preempt. To avoid mind numbing repeats, and in keeping with our *'as simple as possible'* maxim, we use a *least-recently-used* policy, with random choice breaking ties.

4 Results

It takes a considerable amount of effort to stabilize a system so that it can be used by random subjects, 'off the street'. Our infrastructure, documented in [12, 13], has achieved this – we did not experience a single system crash. Moreover, the system was not patched over the period of our trials. To date we have conducted 5 trials with the human wizard controller and 10 trials with the ASAP controller. All position information, utterances, (and image and audio streams for Wizard trials) are logged with time-stamps to our database. All trials have used the same phone (Sony Xperia Acro S), data plan (Telia 3G), Google's TTS engine, the same back-end servers, database, etc.

Figure 2 gives a qualitative overview of our results. The trajectory in black, labeled 'Ideal' shows an upper bound on the effectiveness of a controller; the 'subject' has perfect knowledge of the tour and walks at a brisk pace. The trajectory in blue, labeled 'Wizard', shows a typical trial of a subject guided by the wizard. Finally the trajectory in green, labeled ASAP, shows a typical trial guided by our ASAP controller. Figure 3 shows a degenerate case for our ASAP controller which we discuss below. We have tabulated the following simple measures over our trials:

		Goals Reached			
Controller	# trials	avg	med	min	max
Ideal	1	7			
Wizard	5	4.6	5	4	5
ASAP	10	4.4	5	1	6

Using NTP we instrumented our system to measure, in real-time, latencies for GPS reports to be transmitted to the controller. Values are highly variable. For example, one of our trials had an average 0.27 sec (max 1.45 sec) latency while another had a 1.30 sec average (max 13.96 sec).

Our exit interviews for ASAP trials were mixed. Some subjects, generally those that performed well (e.g. reaching five or six goals), were very enthusiastic saying that they 'want this app!' Others, especially those that did not score well, complained of receiving instructions at poor times, receiving contradictory instructions, or being asked to do the impossible (e.g. walk through bushes). Some subjects expressed annoyance of exact meter reports ("turn right in 34 meters."). Some were annoyed by the 'military style' brevity and constant oversight, however others liked that style. Many mentioned that the references to landmarks made the experience much more satisfying. All found the TTS perfectly intelligible.

5 Discussion

Our goal with the ASAP controller was to engineer a baseline pedestrian guidance system that could, albeit perhaps inefficiently, guide subjects along routes. We developed earlier, simpler controllers that failed at this; such controllers routinely put subjects into bad control loops similar to that of figure 3. The distinguishing ingredient in the ASAP controller is simple time series analysis of variability of measures over the last 5-10 seconds of the subjects state. Only when

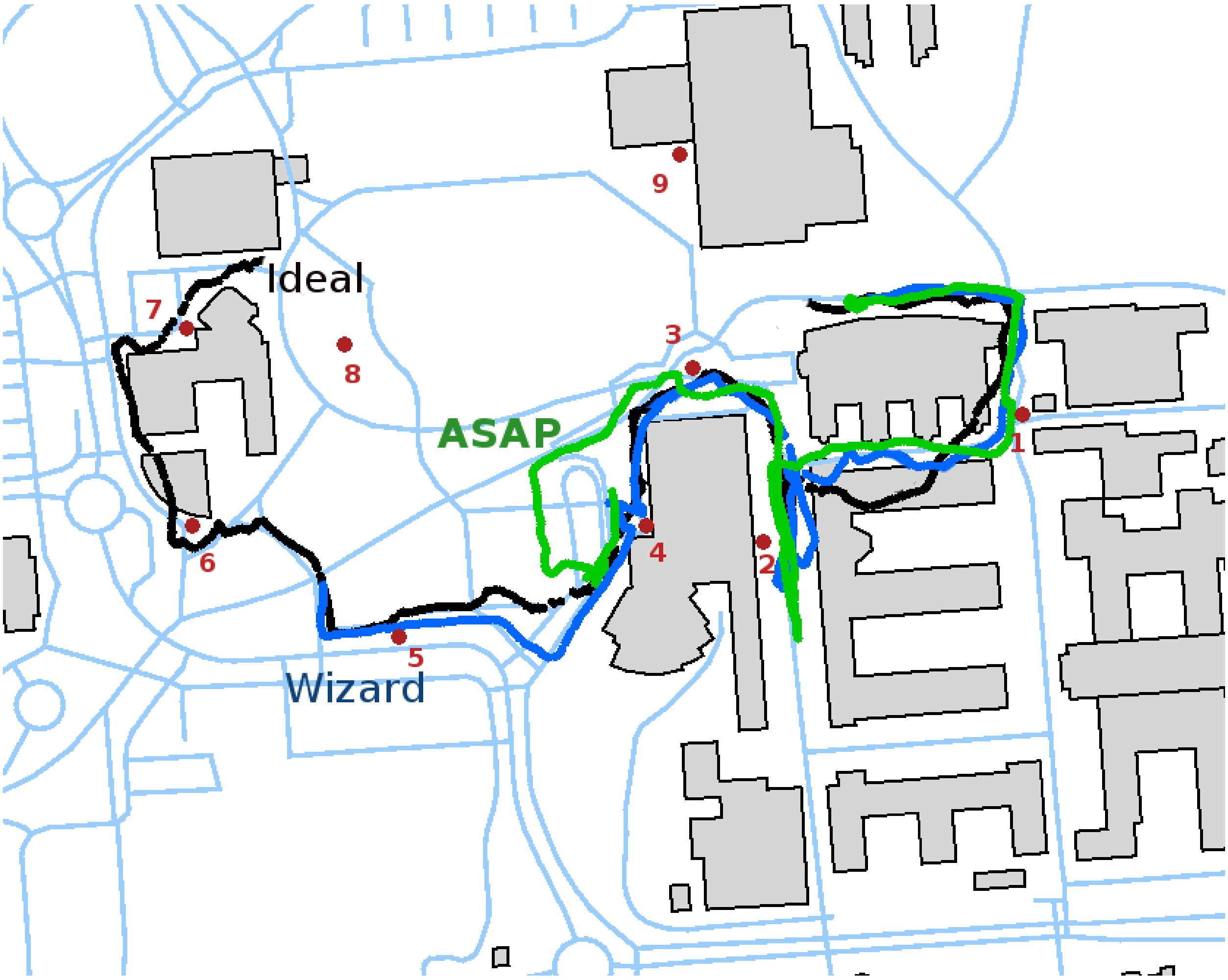

Figure 2: Typical trajectories

variability is below certain thresholds, are certain speech acts tolerated. Without such checks, it is hard to imagine a pedestrian navigation system succeeding.

While our evaluation is of a small, qualitative nature, we argue that it provides some evidence that TTS-based navigation is feasible; the control channel is rich and stable enough to adequately guide subjects. The human wizard appears to be slightly more effective on average and certainly more reliable than the ASAP controller. Both controllers are at least approaching the ideal. Interestingly the best performing trials were under the ASAP controller. Our hunch is that the ASAP controller, tireless as it is, out performs a human wizard so long as map incompleteness, GPS errors, large latencies or other unknown conditions, don't force it into a degenerate case.

Figure 3 shows degenerate case of the ASAP controller (our only trial where only one goal was reached). In this trial a report about reaching the first goal, caused the user to stop abruptly. The user then failed to have a heading and was asked to go forward. Because of the confined space and simultaneously a period of increased network latency, the system and subject entered into a

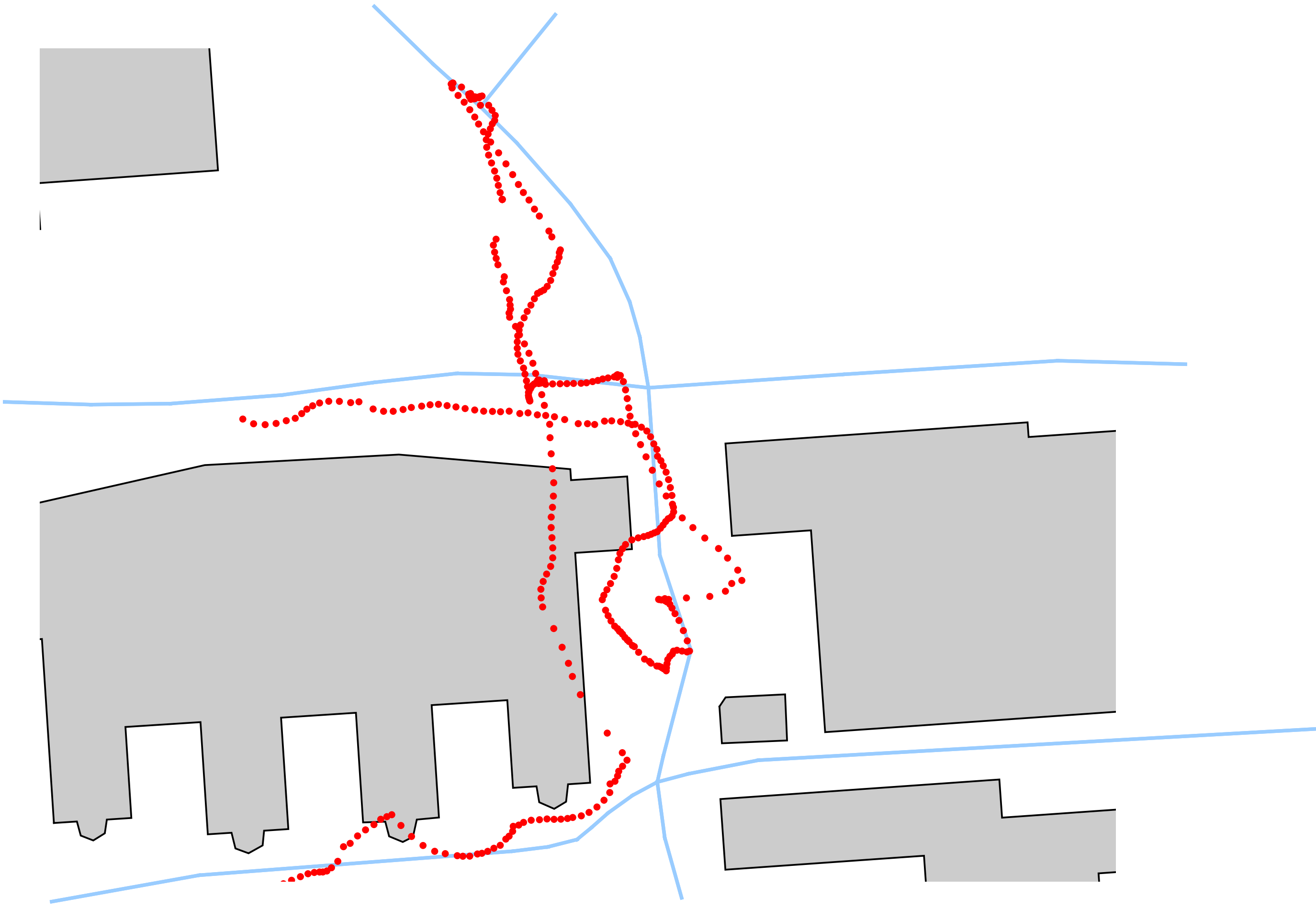

Figure 3: Degenerate case for ASAP controller

loop where the subject moved back and forth erratically, and the system gave poorly coordinated, overcompensating instructions. Some, but not all, of the other ASAP trials had similar control instabilities, although less dramatic. Future controllers will endeavor to recognize and avoid such degenerate control loops, perhaps falling back to more robust, though inefficient strategies.

6 Conclusions

In this report we provide some evidence that TTS based navigation can be made to be practical. To strengthen these conclusions, we must replicate these findings in more varied locations and routes, over larger, more varied populations. We will implement more advanced controllers, especially those that seek to predict pedestrian position and schedule utterances [11]. We will also test the effect that simple confirmation dialogues [2] have toward improving efficacy, perhaps comparing rule-based versus statistical [8] approaches. We welcome other groups to test their approaches over this infrastructure and/or experimental protocol.

All of the data collected for this report will accompany the open-source release of our software, soon available for public downloaded at `http://janus-system.eu`.

References

[1] P. Bartie and W. Mackaness. Development of a speech-based augmented reality system to support exploration of cityscape. *Transactions in GIS*, 10(1):63–86, 2006.

[2] J. Boye, M. Fredriksson, J. Götze, J. Gustafson, and J. Königsmann. Walk this way: Spatial grounding for city exploration. In *Proc. 4th international workshop on spoken dialogue systems, IWSDS'2012*, Paris, France, November 2012.

[3] S. Coast. How OpenStreetMap is changing the world. In *proc. of W2GIS*, page 4, 2011.

[4] R. Dale, S. Geldof, and J.-P. Prost. Using natural language generation in automatic route description. *Journal of Research and Practice in Information Technology*, 37(1), 2005.

[5] M. Dräger and A. Koller. Generation of landmark-based navigation instructions from open-source data. In *EACL*, pages 757–766, 2012.

[6] S. Janarthanam and O. Lemon. *The GRUVE challenge: generating routes under uncertainty in virtual environments*, pages 208–211. 2011.

[7] S. Janarthanam, O. Lemon, P. J. Bartie, T. Dalmas, A. Dickinson, X. Liu, W. A. Mackaness, and B. L. Webber. Evaluating a city exploration dialogue system with integrated question-answering and pedestrian navigation. In *ACL*, pages 1660–1668, 2013.

[8] S. Janarthanam, O. Lemon, and X. Liu. A web-based evaluation framework for spatial instruction-giving systems. In *ACL (System Demonstrations)*, pages 49–54, 2012.

[9] S. Janarthanam, O. Lemon, X. Liu, P. J. Bartie, W. A. Mackaness, T. Dalmas, and J. Goetze. Integrating location, visibility, and question-answering in a spoken dialogue system for pedestrian city exploration. In *SIGDIAL Conference*, pages 134–136, 2012.

[10] A. Koller, K. Striegnitz, D. Byron, J. Cassell, R. Dale, J. D. Moore, and J. Oberlander. The first challenge on generating instructions in virtual environments. In *Empirical Methods in Natural Language Generation*, pages 328–352, 2010.

[11] M. Minock and J. Mollevik. Prediction and scheduling in navigation systems. In *Proceedings of the Geographic Human-Computer Interaction (GeoHCI) workshop at CHI*, April 2013.

[12] M. Minock, J. Mollevik, and M. Åsander. Towards an active database platform for guiding urban pedestrians. Technical Report UMINF-12.18, Umeå University, October 2012.

[13] M. Minock, J. Mollevik, M. Åsander, and M. Karlsson. A test-bed for text-to-speech-based pedestrian navigation systems. In *Proc. of the 18th International Conference on Applications of Natural Language to Information Systems (NLDB)*. Springer Verlag, 2013.

[14] M. Pielot, B. Poppinga, W. Heuten, and S. Boll. PocketNavigator: studying tactile navigation systems in-situ. pages 3131–3140, 2012.

[15] M. Raubal and S. Winter. Enriching wayfinding instructions with local landmarks. In *GIScience*, pages 243–259, 2002.

[16] K.-F. Richter and A. Klippel. A model for context-specific route directions. In *Spatial Cognition*, pages 58–78, 2004.

[17] A. Vlachos, S. Clark, T. Dalmas, R. Hill, S. Janarthanam, O. Lemon, M. Minock, and B. Webber. D4.1.2: Final request analysis component. Technical report, SpaceBook project technical report, 2014.