SMALLWorlds – a Multi-lingual Speech Corpus for Cognitive Research

Peter Juel Henrichsen, Marcus Uneson

Center for Computational Modelling of Language, Centre for Languages and Literature,
Copenhagen Business School,Lund University
pjh.isv@cbs.dk,marcus.uneson@ling.lu.se

Abstract

We present the speech corpus *SMALLWorlds* (Spoken Multi-lingual Accounts of Logically Limited Worlds), newly established and still growing. SMALLWorlds contains monologic descriptions of scenes or worlds which are simple enough to be formally describable. The descriptions are instances of *content-controlled monologue*: semantically "pre-specified" but still bearing most hallmarks of spontaneous speech (hesitations and filled pauses, relaxed syntax, repetitions, self-corrections, incomplete constituents, irrelevant or redundant information, etc.) as well as idiosyncratic speaker traits. In the paper, we discuss the pros and cons of data so elicited. Following that, we present a typical SMALLWorlds task: the description of a simple drawing with differently coloured circles, squares, and triangles, with no hints given as to which description strategy or language style to use. We conclude with an example on how SMALLWorlds may be used: unsupervised lexical learning from phonetic transcription. At the time of writing, SMALLWorlds consists of more than 250 recordings in a wide range of typologically diverse languages from many parts of the world, some unwritten and endangered.

Keywords: content-controlled monologue, semantic unit selection, spatial grid networks, multilingual speech corpus

1. Introduction

Most linguistic corpora, whatever other dimensions they may keep constant, collect data with no restriction on semantics (except perhaps what the domain may suggest, on a general level). To be sure, this is very reasonable. The ultimate goal of corpus processing is often to build a system which can make statements about the meaning of unknown language samples – clearly, with that aim, semantics cannot be treated as a given.

Nevertheless, there are interesting applications, and also research methodologies, which make good use of exhaustively pre-specified semantics. More specifically, they may take as point of departure a formal description F of a scene S and (some representation of) a natural-language description L of the same scene, and exploit (or explore) mappings between linguistic labels in L and meanings in F. Sometimes finding the mappings themselves is the main interest; sometimes taking them as given help in answering questions about one linguistic aspect or another of L.

In this paper, we present the growing corpus *SMALLWorlds* (Spoken Multilingual Accounts of Logically Limited Worlds), built on this premise. "Exhaustively specified semantics" is admittedly a mouthful, and it should be clear that the worlds we refer to are toy-sized – indeed, "logically limited". However, along with the details of the corpus, we will give an example of an interesting question one might seek to answer using such a data set, and we will leave several other suggestions for the future.

Corpora containing natural-language descriptions of such formally describable microcosmoses are unusual, but not entirely unheard of. In the early eighties, Levelt let Dutch speakers describe a number of "spatial grid networks", simple figures consisting of differently coloured dots connected by lines on a grid, to investigate aspects of the linearization order in self-paced monologues (Levelt, 1989). Swerts and Collier (Swerts and Collier, 1992) used a slightly more complex network to study prosodic correlates of discourse units in spontaneous speech, again eliciting self-paced monologues. Rather than the prosodic findings, their main point is actually the method *per se*: given a network like that in Figure 1, the linearization principles suggested by Levelt (Levelt, 1989) enable experimenters to predict the speakers' path through it, their chaining of statements, and the predicative content of these statements; and by judiciously contrasting shape or colour or both between current and next shape, experimenters gain control over new and given information at any point, with only very general instructions given. The same or a similar setup is well suited for testing other hypotheses on spontaneous speech – say, phonetic, syntactic, or lexical correlates of the distinction between new and given, or of that between content and function words (Swerts and Collier, 1992).

More generally, the elicitation of such descriptive language is, in the words of Swerts and Collier, "characterized by the fact that a speaker is constrained by the experimenter with respect to what he will say, while keeping his speech output spontaneous." As they note, this is a powerful and versatile experimental paradigm. In such *content-controlled monologue*, the experimenter chooses by careful stimulus design the expected predicates and their order, leaving choicepoints only when these are part of the experiment. The task is well-defined and practicable and the subject knows exactly what information to convey. Yet, the instructions can be minimal and no words need to be put in the speaker's mouth. The resulting descriptions will bear many or most hallmarks of truly spontaneous speech (at least as far as monologues go): hesitations and filled pauses, relaxed syntax, repetitions, self-corrections, incomplete constituents, irrelevant or redundant information, etc. Most characteristics of a particular speaker's natural style will also be preserved, as will temporal aspects of spontaneous speech (e.g., control switching between planning and execution).

To these observations on content-controlled monologue (Swerts and Collier, 1992), we may add a few of our own.

First, the descriptions refer to scenes which are easily formalizable. This is highly desirable for any automatic processing. Second, the scenes are composed of abstract, geometric shapes, carrying few or no extra-figural connotations and associations,[1] and thus do not encourage idiosyncratic digressions. Third (somewhat related to the previous two), the scenes are simple enough that any description in human language need only use a small number of content morphemes (perhaps 20-30). The small vocabulary involved makes studies on even a single description feasible (and rewarding). Fourth, the scenes are reasonably culture-independent; they should be describable in any human language using only a small set of concepts (but see also Section 2.3.).

Subsections of SMALLWorlds have been used in research reported elsewhere (Henrichsen, 2011; Henrichsen and Christiansen, 2011). However, the current paper is the first one to present the corpus in its entirety, and under that name.

2. The SMALLWorlds Corpus

In this section, we describe the details of SMALLWorlds as of January 2012. More experimental ideas are not covered here, although we mention a few of them in Section 4..

2.1. Technical details

The majority (around 210) of the SMALLWorlds recordings were made by Uneson in 2010. These were all recorded with a Neumann U87 microphone in the anechoic chamber of the Humanities Lab at Centre for Languages and Literature, Lund University. The rest of the recordings were made by Henrichsen in a more experimental spirit, with irregular intervals during 2005-2011. These were made with portable high-quality sound equipment in quiet surroundings. Most are in mono, some in stereo (two microphones, one stationary and one chin-mounted). All data files, whether recorded in the field or not, are stored as wav-files, sampled at 44.1kHz, 16 bit or better. Currently, there are about nine hours worth of recordings (minimal duration 43", median 137", max 614"). Video files are available for a few sessions.

2.2. Participants

We deliberately designed the task set of the SMALLWorlds sessions to be lightweight enough (a typical session lasting around 20 minutes) that it would be feasible to find volunteers without necessarily having made previous appointments. This strategy has worked well. Thanks to the light demands on both informant and instructor, we have been able to set up recording sessions with little or no preplanning: during coffee breaks, in waiting rooms, between lessons, etc. In addition, we have been able to record also in remote mountain villages and sparsely populated rural areas.

For the Lund recordings, the vast majority of the participants were international students aged 20-30. The rest of the participants have a larger spread. With few exceptions, speakers were recorded in their first language (or in one of them, in cases of multilinguals with self-assessed native competence in more than one language). In addition, a few self-assessed bilinguals performed the task in both languages.

2.3. Tasks and stimuli

All SMALLWorlds sessions referred to in this section contain a description of the network in Figure 1 (or, in early sessions, a hand-drawn equivalent).[2] This particular network was originally used in the DanPASS corpus (Grønnum, 2009) (and thus these subtasks are compatible in the two corpora). The informants received oral instructions along these lines: "Please describe the drawing, so that it can be later reconstructed according to your description. You should mention all the coloured objects, and also their distribution on the paper. Begin with the figure to which the arrow points".

On increasing the typological coverage, some cross-linguistic issues arose. For instance, the well-known linguistic fact (Berlin and Kay, 1969) that not all languages have lexicalized labels (monolexemic or not) to cover all the colour contrasts became apparent. Thus, 'red' and 'brown' may in most contexts translate to the same lexeme, just as Russian голубой *goluboy* 'light blue' and синий *siniy* '(dark) blue' are often rendered simply as 'blue' in English. Similarly, in languages with little use in educational settings, there may be no names for mathematical concepts such as circles or squares.

This is not really a problem in a functional sense – what we require of the labels is that they solve the task at hand, not that they are monolexemic, known to all speakers, or non-loans. Thus, the speakers were encouraged to solve the problems as they would in ordinary conversation – perhaps by a loan from another language which could be assumed familiar, or by a paraphrase such as "coffee-coloured" for brown.

Nevertheless, in later sessions, we constructed an additional network with target words chosen from the Swadesh list (Swadesh, 1971), a set of culture-independent concepts which can be expected to have labels in all languages (*sun, water, ear*, etc.). The shapes chosen to represent these meanings were intended to be readily recognizable prototypes, stylized enough not to elicit irrelevant details (Figure 2). At the time of writing, we have about 140 descriptions of the Swadesh network.

2.4. Composition of the corpus

An overview of the corpus composition of SMALLWorlds is given in Table 1. For a set of selected languages (more or less the ones in Table 1), our aim is to get 15-20 speakers, roughly balanced for sex, which should be enough for most tasks where inter-speaker comparisons are important.

This particular selection of prioritized languages is of course partly a convenience sample, a compromise between typological spread and expected availability of subjects, annotation expertise, and other external resources. In parallel, however, we have also followed the "rare butterfly" policy of recording a few speakers of all languages we could get

[1] Even hardcore cat or dog persons will have a hard time projecting strong likings or aversions onto a blue square.

[2] When use as stimulus, the figures were printed on paper and enlarged to roughly 15x20 cm.

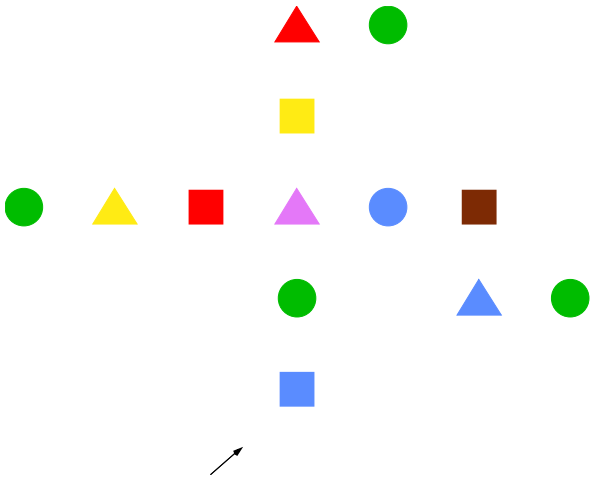

Figure 1: The geometrical network

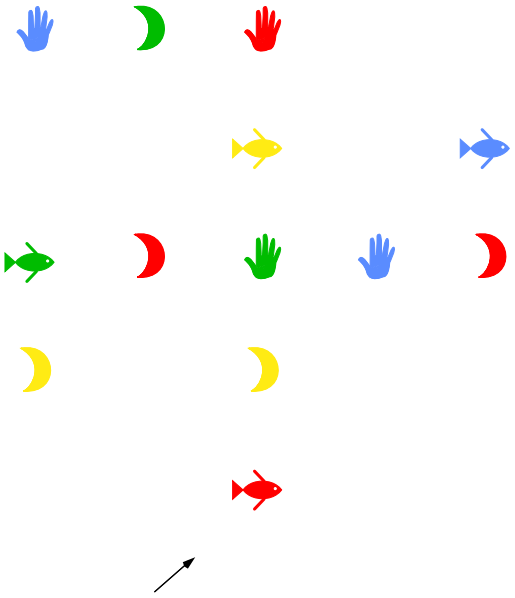

Figure 2: The Swadesh network

hold of. Many of these recordings may for practical reasons never be annotated, but they still make interesting comparison. As we argue above, several interesting questions can be asked from a single description.

Finally, and along more experimental dimensions, we have included recordings by children (aged 6-9), speech impaired adults, L2 speakers of badly broken English, and illiterates. We even sport a whispering gallery of sessions with voiceless phonation, and occasional recordings in semi-shouted style. These deviant styles represent challenges that are often ignored or under-estimated in, for instance, speech technology projects.

Language	F	M	Total
English	9	10	19
Swedish	9	10	19
German	9	8	17
Japanese	8	9	17
Spanish	8	8	16
Danish	9	5	14
Chinese	9	2	11
Finnish	8	3	11
Persian	5	6	11
Hindi	7	4	11
Tamil	4	4	8
French	3	5	8
Arabic	2	6	8
Russian	6	0	6
Italian	3	3	6
Dutch	1	4	5
Turkish	5	0	5
Korean	1	4	5
Other (36)	36	25	61
Total	142	116	258

Table 1: SMALLWorlds subjects, as of January 2012. "Other", with 4 speakers or less, includes Albanian, Azeri, Bengali, Bosnian, Bulgarian, Cantonese, Catalan, Croatian, Czech, Flemish, Greek, Hebrew, Hungarian, Icelandic, Igbo, Irish, Kammu, Khmer, Latvian, Lithuanian, Luxemburgisch, Nepali, Newari, Norwegian, Polish, Portuguese, Pulaar, Punjabi, Romanian, Swiss German, Thai, Toda, Ukrainian, Urdu, Vietnamese, Wu Chinese

2.5. Annotation

SMALLWorlds is available as Praat TextGrids.[3] The annotation depth is quite heterogeneous. At the extremes, some recordings are very fully annotated (syllable-level time coding, orthographic and phonetic transcription, fine-grained PoS, acoustic measurements, prosodic features, meta-linguistic commentary, detailed background information) whereas most of the third-world recordings are much more sparsely annotated (coarse time-coding, lexical lists of colour/shape/location/direction terms, and meta-linguistic commentary only).

The majority of the recordings, including those in English, German, Spanish, Swedish, Dutch, Finnish, Arabic, French, and Turkish (Table 1), are somewhere in between, typically including pause-based time coding, orthographic (or pseudo-orthographic) transcription, acoustic measurements, and some meta-linguistic information.

For every recording, irrespective of annotation level, we have been very particular about registering the naming sequence, i.e. the order in which the coloured objects were introduced by the informant. Also the central lexical terms (in the task shown, the names of the colours, shapes, spatial relations, and directions) are on file in almost all cases.

[3] http://www.fon.hum.uva.nl/praat/

sitten mennään takaisin tähän ylöspäin
then we go back to this upward
suuntautuvaan pylvääseen ja tuohon lilanväriseen
oriented column and to that lilac-colored
kolmioon siitä oikealle on sininen pallo ja
triangle of that to the right is blue ball and
sitten ruskea neliö
then brown square

'Then we go back to the vertical column and the lilac triangle. To the right of that, there is a blue ball and then a brown square.'

eh ahora volvemos atrás al triángulo de
eh now we return back to the triangle of
color rosa del que hablábamos
colour pink about which we were talking
anteriormente donde se se producía una
earlier where there there was a
trifurcación y y en esta ocasión vamos a
three-way junction and and this time we will
coger el camino que va hacia la derecha
choose the path which goes to the right
según miramos según andamos en ese camino eh
as we see it as we go on this path eh
hay en realidad cuatro figuras eh dos de
there is really four shapes eh two of
ellas siguen la misma línea y después el
them continue the same line and then the
camino se ca- cambia de dirección bueno la
path ch- change direction okay the
primera figura que encontramos es un círculo de
first shape that we meet is a circle of
color azul que está unido a un cuadrado
colour blue which is connected to a square
de color marrón
of colour brown

Figure 3: Orthographic transcription (time codes omitted) of two corresponding excerpts from SMALLWorlds descriptions: Finnish (top) and Spanish.

In addition, there are semantic annotations as outlined in Section 2.6. for one or two speakers for each of English, German, Spanish, and Swedish.
Of course, annotation work is still in progress (depending crucially on available funding). Future annotation efforts will concentrate primarily on the prioritized languages.
Two excerpts in orthographic transcription, one Finnish and one Spanish, are shown in Figure 3. They have the same semantic content: both describe the central pink triangle and the blue circle and the brown square to the right of it, Note the presence of speaker characteristics, in particular the large variation in verbosity.

2.6. Recombined data: Semantic unit selection

We have developed a technique for building new network descriptions out of existing ones, such that these new descriptions have a certain, user-defined amount of redundancy, self-corrections, etc. – a crude but useful way of parameterizing speaker personality. The method, which we have termed *semantic unit selection*, has been developed specifically for manipulating the data of SMALLWorlds; in particular, we have used it to provide training and evaluation data at controlled levels of difficulty for inference-based learning (cf. Section 3.).
Semantic unit selection is based on recombination of utterances, where we define utterances to be phonetically independent carriers of meaning with respect to some given domain. The utterances to be recombined are selected based on their semantic content. The method is not dependent on the original world described but can take a formal description of a new world as input, or even enumerate all worlds describable by the data (along with their respective descriptions, in formal and synthesized natural language, in whatever representation the corpus employs).
The limitations of the method must be made clear from the start: it is useful only for restricted, formally describable domains, and – to avoid any possible misunderstandings – it is not a way of attacking data sparseness problems. It will certainly not be able to pull data which was not present in the original corpus out of some magic hat. Still, it has served us well as a way of preparing or filtering data with a controlled degree of difficulty for learning and evaluation tasks.
We describe semantic unit selection in more detail elsewhere (Uneson and Henrichsen, 2011), but very briefly, it works as follows. First, the data is segmented into units which are large enough to be rearrangeable without regards to local phonetic phenomena, such as coarticulation and assimilation – an automatic segmentation method based on crude measures such as pause durations (say, 200 ms) works well enough for this purpose. Next, each utterance is annotated with respect to its semantic content, expressed in first-order predicate logic; specifically, the current implementation uses the Definite Clause Grammar (DCG) formalism of the Prolog language (Pereira and Warren, 1980). Finally, in a generating step, utterances are selected top-down and compositionally combined to larger units – a string of segments in whatever representation the terminals stand for. The selection can be guided, so that (say) concise units are preferred to wordier ones or vice versa – this mechanism offers a crude view on speaker personality.

3. Learning lexical structure

As a concrete usage example, we present an experiment with unsupervised lexical learning from unsegmented phonetic transcription. The goal of this experiment was to design a robust learning device able to determine the basic vocabulary (the shape and colour terms) and the path (the naming sequence) of a description session. The learning device had access to two information sources only, the transcription data (unsegmented, i.e. without word segmentation) and a minimal formal model of the "small world" of Figure 1. The formal model was represented as a compact

```
prop(colour,blue,[e1,e10,e12]).
prop(colour,green,[e2,e6,e9,e13]).
prop(colour,red,[e4,e8]).
prop(colour,yellow,[e5,e7]).
prop(colour,purple,[e3]).
prop(colour,brown,[e11]).
prop(shape,square,[e1,e4,e7,e11]).
prop(shape,circle,[e2,e6,e9,e10,e13]).
prop(shape,triangle,[e3,e5,e8,e12]).
```

Figure 4: A formal model of the colours and shapes of the network in Figure 1, expressed as Prolog clauses.

set of logical facts, shown in Figure 4 as clauses in the programming language Prolog.

In this experiment, object names `e1` to `e13` were used for the thirteen coloured objects, `e1` referring to the lower blue square, and `e2 ... e13` applied "clockwise". Observe how the colour "blue" is represented semantically as in first order models of classical predicate logic, viz. as the complete set of objects `[e1,e10,e12]` sharing that property. Apart from the formal model, the learning device did not contain any lexical, morphological or semantic categories, in short, no language knowledge. The main inference engine had to figure out which lexical mappings of the continuous speech stream (represented by the unsegmented transcription) could match a well-formed description of the illustration (represented by the formal model).

When used as experimental data, the string of phones (excluding marking of stress and vowel length) is processed in three stages:

1. all frequently occurring n-grams, such as [ililinjɪ], are identified;
2. the n-grams are arranged in sets of three based on distributional similarity, such as [[f i ʁ k a], [i l i l i n j ɪ], [t ʁ æ k a n d]];
3. the triplets are piped to the inference module as lexical hypotheses.

For identification and arrangement of n-grams (steps 1 and 2), we employed the algorithm Siblings & Cousins (Henrichsen, 2004). This algorithm exploits the fact that two words with complementary semantics (like two distinct colour terms, say *blue* and *green*) tend to prefer similar contexts at their right and left edges. For instance, consider a pair of corpus instances *a blue circle* and *a green circle*, both quite frequent in the descriptions. In this case, of course the colour terms share the context *a _ circle*. Quantifying over all n-gram candidates and all their respective context functions, the Siblings & Cousins algorithm produces analyses as the one shown in Figure 5.

Based on their left and right context selection functions, the n-grams [t ʁ æ k a n ʔ d] and [t ʁ æ k a n ʔ d] are thus – not surprisingly – judged to be similar to a degree of 100%. More interestingly, the n-grams [t ʁ æ k a n ʔ d] and [f i ʁ k a n ʔ d] score 83.7%, meaning that these two n-grams do indeed prefer the same contexts to a high degree. This is satisfactory, the two n-grams representing the Danish words for triangle and square, respectively. As can be seen, some semantically neutral variation in pronunciation is also detected (e.g. the t/d allophones).

```
#93 n-gram analysed: [t ʁ æ k a n ʔ d]
1.000000 [t ʁ æ k a n ʔ d]
0.837132 [f i ʁ k a n ʔ d]
0.727861 [l e l a t ʁ æ k a n ʔ d]
0.646050 [s i ʁ g l]
0.629778 [t ʁ æ k a n ʔ t]
0.625563 [d ɛ n ʔ]
0.614339 [f i ʁ k a n ʔ t]
```

Figure 5: Sample from Siblings & Cousins log (stage 1 and 2, see text). Analysed n-gram #93: [t ʁ æ k a n ʔ d] (the Danish word for triangle). Listed: high-scoring n-grams, sorted by context selectional similarity with [t ʁ æ k a n ʔ d]. Male speaker, session ID m_014g

```
triangle: [t ʁ æ k a n]
square: [f i ʁ k a n]
circle: [s i ʁ g l̩]

blue: [b l ɔ]
green: [g ʁ œ n]
red: [e n ʁ œ ð]
yellow: [g u l]
brown : [s,ɔ,a,d,ʊ,e,n,b,r,u,n]
purple : [a,d,ʊ,n,l,e,l,a]
PATH : [e1,e2,e3,e10,e11,e12,e13,e4,e5,e6,e7,e8,e9]
```

Figure 6: Sample from learning log (stage 3, see text; cf. also Fig. 5), showing the deduced lexemes and the associated path

Note in Figure 5 the n-gram [l e l a t ʁ æ k a n ʔ d]. This n-gram corresponds to a compound expression (purple triangle), but was nevertheless picked by the algorithm as a possible semantic atom (based on its contextual similarity with a shape term proper). This judgment is actually not surprising, given the fact that colour [lela] 'purple', is represented (Figure 1) by one object only, creating a strong cohesion effect. This interplay between atomic and compositional semantic readings of compound words is of course well-known in human discourse too: the term *red herring* may occasionally be used to refer to a herring which happens to be red; but usually its meaning is atomic.

Turning now to step 3: for each triplet, the inference engine searches for a division of the entire transcription into 13 subsections (corresponding to the 13 objects in Figure 1), each containing a triplet element (the one in the example would thus be rejected, [ililinjɪ] not being a shape name). On success, the 13-section is checked for consistency with human description strategies, and a corresponding colour mapping is deduced (Figure 6).

The particular learning session that the figures refer to used data from a recording from 2011, of an eight-year old Danish boy (dk008). The path was correctly identified by the learning device: informant dk008 did name the thirteen objects in the order shown. Concerning the deduced vocabulary, several unusual phonetic forms are encountered. Perhaps most surprising are the very long terms derived for colours *brown* and *purple* in comparison to their dictionary forms [b,r,u,n] and [l,e,l,a]. With a bit of reflection, though, it is easy to understand why the inference engine, with each of these colours occuring only once in the described picture, has too sparse data to determine their standard lexical forms. Since the learning algorithm build on a more-is-better strategy, the inference engine instead expands the terms as far as the logical constaints allow, arriving at the unusual, but entirely natural conclusion. In the same vein, *red* translated to [e,n,r,œ,ð] rather than the expected form [rœð] since the latter, in all its occurrences in the dk008 session, is preceded by [e,n]. Across all Danish learning sessions we have performed, we have encountered five different renderings of colour yellow: [e,n,g,u,l], [n,g,u,l], [g,u,l], [ŋ,g,u,l], [d,ɔ,e,n,g,u,l]; the dictionary form [g,u,l] is not even the most frequent one.

For a larger-scale verification of this result, we borrowed 18 of the sessions from the corresponding subtask of the DanPASS corpus (Grønnum, 2009), which as pointed out above is compatible with the SMALLWorlds corpus. Very briefly, the learning experiments for this larger set (n=22) were successful in the sense that all results reported by the learning device were in agreement with the human judgments. The central lexemes (colour and shape terms) were thus correctly identified (allowing skewed delimitations) in 16 out of 22 sessions. The remaining 6 cases all contained factual errors (e.g. informants referring to a yellow square as a "yellow triangle", or failing to specify the colour of one of the objects). In other words, for this subcorpus the learning device was 100% successful with regards to the experimental premises. Further details on the practical and computational installments are found in Henrichsen (2011).

4. Future directions

The SMALLWorlds corpus presented here is an early version – practically all informants have faced tasks identical or very similar to the one in Figure 1. Currently, we are exploring new dimensions of this basic design, allowing for controlled variations of the complexity of the stimulus while keeping the main idea of exhaustively describable semantics. For instance, we aim to present individual informants with a series of description tasks, varying systematically the diversity of colours, shapes, and spatial relations. Such description data will allow us to study in more detail the relations between the formal properties of the scene described and the typological, individual, and situational variation. Another interesting variation involves using shapes with inherent orientation, such as stylized images of cars, buildings, or trees, to study the effects on perspective chosen.

We are also considering on-screen networks which evolve dynamically, interactively steered by the informant's mouse-clicks. On a more dialogical note, we are taking the first steps towards an experimental design with a two-way spoken interface between man (informant) and machine (inference engine). In the not so distant future, we intend to perform interactive dialogue-style experiments where terms and relations are negotiated on-the-fly rather than inferred *post festum*.

Turning to usage rather than content, we conclude with a non-exhaustive list of suggestions for linguistic (or cross-linguistic, where appropriate) explorations, which we believe the data lends itself to:

- speech planning and linearization (cf. (Levelt, 1989; Levelt, 1982b; Levelt, 1996))
- cognitive styles in spatial descriptions (cf. (Levelt, 1982a))
- situational, individual and typological variation in choice of frame-of-reference (cf. (Levinson, 2003))
- inference-based learning of lexical items (cf. (Henrichsen, 2011))
- prosodic modulation as a marker of information structure (cf. (Swerts and Collier, 1992; Swerts, 1994; Henrichsen and Christiansen, 2011))
- relations between rhetorical structure and information structure

Even more experimentally, we believe the data can be a good companion when setting foot on less traveled roads: unsupervised models for machine translation in restricted domains; self-learning speech recognition in restricted domains; early-stage L1 acquisition models.

We will soon establish a dedicated website for the SMALLWorlds corpus including contact and download information, terms of use of the corpus data, references to dedicated and secondary literature, links to the test materials, and more. Just google "SMALLWorlds" to keep updated.

5. Acknowledgements

Marcus Uneson wishes to acknowledge the support of the Humanities Lab at Centre for Languages and Literature, Lund University.

6. References

Brent Berlin and Paul Kay. 1969. *Basic color terms: Their universality and evolution*. Univ of California Press.

Nina Grønnum. 2009. A Danish phonetically annotated spontaneous speech corpus (DanPASS). *Speech Communication*, 51:594–603.

Peter Juel Henrichsen and Thomas Ulrich Christiansen. 2011. Information based speech transduction. In *Proceedings of International Symposium on Auditory and Audiological Research, ISAAR 2011*.

Peter Juel Henrichsen. 2004. Siblings and cousins: statistical methods for spoken language analysis. *Acta Linguistica Hafniensia*, 36:7–33.

Peter Juel Henrichsen. 2011. Fishing in a speech stream, angling for a lexicon. In *NODALIDA-2011*, May. Riga University.

Willem Levelt. 1982a. Cognitive styles in the use of spatial direction terms. *Speech, place, and action*, pages 251–268.

Willem Levelt. 1982b. Linearization in describing spatial networks. pages 199–220. D. Reidel Dordrecht,, The Netherlands.

Willem Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Willem Levelt. 1996. Perspective taking and ellipsis in spatial descriptions. In P. Bloom, M. Peterson, L. Nadel, and M. Garrett, editors, *Language and space*, pages 77–108. Cambridge, MA: MIT Press.

Stephen Levinson. 2003. *Space in language and cognition*. Number 5 in Language, culture, and cognition. Cambridge University Press.

Fernando. Pereira and David Warren. 1980. Definite clause grammars for language analysis–a survey of the formalism and a comparison with augmented transition networks. *Artificial intelligence*, 13(3):231–278.

Morris Swadesh. 1971. *The origin and diversification of language*. Chicago: Aldine. Contains final Swadesh 100-word list.

Marc Swerts and René Collier. 1992. On the controlled elicitation of spontaneous speech. *Speech Communication*, Volume 11(4-5):463–468.

Marc Swerts. 1994. *Prosodic features of discourse units*. Ph.D. thesis.

Marcus Uneson and Peter Juel Henrichsen. 2011. Expanding a corpus of closed-world descriptions by semantic unit selection. In K. Jassem, P. Fuglewicz, M. Piasecki, and A. Przepiórkowski, editors, *Proceedings of Computational Linguistics – Applications (CLA11)*, pages 93–98.