DanPASS – A Danish Phonetically Annotated Spontaneous Speech Corpus

Nina Grønnum
Linguistics Laboratory, Department of Nordic Studies and Linguistics, University of Copenhagen
84 Njalsgade, DK-2300 Copenhagen, Denmark
E-mail: ninag @ hum.ku.dk

Abstract

A corpus is described consisting of non-scripted monologues and dialogues, recorded by 27 speakers, comprising a total of about 70.000 words, corresponding to well over 10 hours of speech. The monologues were recorded as one-way communication with blind partner where the speaker performed three different tasks: (S)he described a network consisting of various geometrical shapes in various colours. (S)he guided the listener through four different routes in a virtual city map. (S)he instructed the listener how to build a house from its individual parts. The dialogues are replicas of the HCRC map tasks (http://www.hcrc.ed.ac.uk/maptask/). Annotation is performed in Praat (http://www.fon.hum.uva.nl/praat/). The sound files are segmented into prosodic phrases, words, and syllables, always to the nearest zero-crossing in the waveform. It is supplied, in seven separate interval tiers, with an orthographical transcription, detailed part-of-speech tags, simplified part-of-speech tags, a phonological transcription, a broad phonetic transcription, the pitch relation between each stressed and post-tonic syllable, the phrasal intonation, and an empty tier for comments.

1. Introduction

Most of our insight into the phonetics of spoken Danish to date is based on carefully manipulated, scripted material read aloud in a sound-treated studio in the laboratory, so-called labspeech. This is not as strange as it may sound to non-phoneticians. First of all, even the largest speech corpora may fail to exhibit a sufficient number of instances of the phenomenon to be investigated, and in the proper context. Secondly, many phonetic phenomena are best studied when the variable under investigation can be carefully controlled and isolated from other, potentially interacting, phenomena. Thus, e.g., the study of tone necessitates control over voicing and aspiration in consonants in the syllable onset and of vowel quality/height; any study of duration calls for control over stress and segmental context; etc., etc. Results obtained from manipulated materials may then serve, at a later stage, as reference for data taken from non-scripted speech. In brief, scripted materials read aloud in the laboratory may lack spontaneity, but they can be made to meet legitimate, specific phonetic research requirements. However, there are a number of very interesting questions about connected fluent speech that cannot be exhaustively answered from samples of read speech. This is especially true of reduction phenomena and of prosody, particularly prosody and its interaction with syntax and pragmatics.

2. Goal

The intention was to supply a corpus for immediate acoustic and perceptual phonetic investigations. I.e. my primary goal is not syntactic, pragmatic, socio-linguistic, psychological, or whichever other aspect of spoken language one might wish to investigate. There are therefore a considerable number of variables that have not been taken into account in the choice of elicitation material. Nevertheless, the corpus may serve as a basis for many linguistic and/or speech technological investigations.

An obvious use is as training material for automatic segmentation and transcription, and it is in fact going to be used for just that purpose in an investigation of acoustic and perceptual building blocks in spontaneously spoken Danish (Christiansen, 2005).

3. The Corpus

The corpus consists of monologues, dialogues and word lists, cf. below. Apart from the word lists, the corpus represents an approximation to speech in a natural setting: The material for elicitation is controlled in the sense that the speakers are given specific tasks to talk about, and they do so in front of a microphone in a recording studio, but their speech is non-scripted.

3.1. Monologues

The monologues were recorded in 1996. They represent various types of instructions. The speaker was seated alone in the recording studio and could communicate with me only via microphone and headphone. Once I had read aloud the instruction for the specific task, cf. below, my audio connection to the speaker was interrupted. In other words, the monologues were recorded in one-way communication with an unseen partner who offered no feedback, neither in the form of questions nor confirmation. Speakers were recorded with – what was then – professional equipment (Sennheiser Microphone ME64, Revox A700, Agfa PEM368 tape).

The speaker first described a network consisting of various geometrical shapes in various colours, cf. Appendix A. It is an elaboration of Swerts and Collier's (1992) network. He or she then guided me through four different routes in a virtual city map, cf. Appendix B, inspired by Swerts (1994). Finally, the speaker – who had a model of the house – told me how to assemble it from its individual parts, cf. Appendix C. This house is an almost exact copy of Terken's (1984) edifice.

3.1.1. Speakers

There were 18 speakers, 13 men and 5 women, all of them students or colleagues in the department, mostly from the greater Copenhagen area and mostly young.

3.2. Dialogues

The dialogues were recorded in the summer of 2004. They are replicas of the Human Communication Research Centre's Map Tasks (Anderson et al., 1991; Brown et al., 1984; http://www.hcrc.ed.ac.uk/maptask/).

The exercise involved the cooperation of two participants. They were seated in separate locations, one a proper professional sound-treated recording studio, the other a recording facility established for the purpose in the control room, damped with curtains of very heavy material surrounding the speaker. They communicated via headsets.

A laboratory set-up like this is hardly the most natural environment for communication, but it turned out to be necessary in order to obtain recordings of sufficiently good quality for subsequent acoustic analysis: Seated in the same room, across from each other with eye-contact, speaker A could invariably be heard through speaker B's microphone, and vice-versa. Whereas I got clean acoustic signals, with no appreciable difference in quality from the studio proper and the ad hoc studio established in the control room. Given the setting, i.e. the lack of visual and direct auditory contact, I assumed that the participants would be most comfortable if they were not also to communicate with a stranger. Accordingly, the two members of a pair knew each other well. They were recorded, via professional headset microphones (Voice Technologies VT700), directly onto CDROMs (HHB Professional Compact Disc Recorder CDR-850) to separate channels in a stereo recording.

Each participant had a map. One, the giver, had a route on his or her map; the other, the follower, did not. Their goal was to collaborate so as to reproduce the giver's route on the follower's map. The maps are not exactly identical, cf. the example in Appendix D: Landmarks are missing on one or the other map, a landmark may appear twice – in two different locations – on one map but not on the other; and the same landmark may have slightly different names on the two maps. This, of course, is what gives rise to a true negotiation, with questions and answers, backtracks, etc. Participants were explicitly informed about these irregularities in written instructions prior to the recording. It was left to them, however, to discover how and where the maps or the designations differed, and to supply the missing items or names on their respective maps. Each pair of speakers completed four different sets of maps, cf. below.

3.2.1. Speakers

22 speakers participated, 13 of whom also recorded the monologues 8 years previously. They are all from the greater Copenhagen area and mostly young, drawn from the pool of (ex-)students and colleagues. There are 13 men and 9 women.

3.3. Word Lists

After completion of the map sessions subjects were asked to read a word list containing all the feature names from the maps they had encountered. Each name appeared twice, in random order, and subjects were asked to read the list in a distinct speech mode. The lists provide citation forms for comparison with the non-scripted dialogue forms. Landmarks and names in the original English version were designed with specific phonological phenomena and processes in mind. I was more or less bound by the nature of the landmarks, with only moderate influence over phonological structure.

3.4. Video Recording

In the studio proper a video-recorder was mounted. The camera was placed as close as possible, and as nearly perpendicular as possible, to the frontal plane of the speaker's face without impeding his/her view of the map. The videos are intended as analysis material for whomever should want to attempt to accompany synthetic Danish speech with a model talking face.

Each speaker had to serve as giver as well as follower, in alternation. Each speaker also had to be video-recorded in both roles. The logistics of running two video-cameras were prohibitive, and we had only one. Accordingly, after two map sessions, with speaker A being giver and follower, respectively, the speakers changed places in order for speaker B to be video-recorded as well. Thus, each pair of speakers had a run through four different sets of maps. A complete recording session lasted 30-40 minutes.

3.5. Statistics

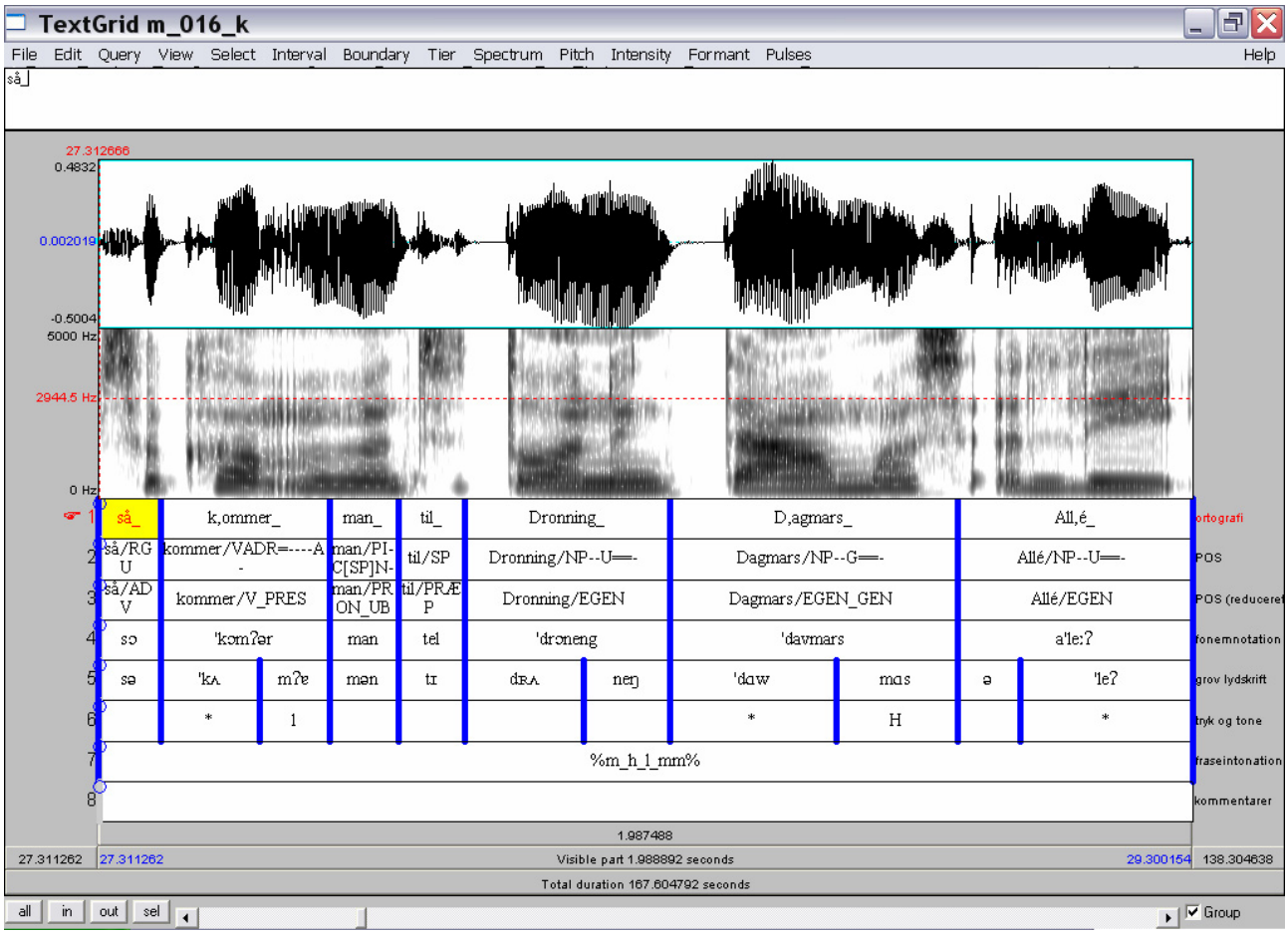
There are well over 10 hours of speech altogether, with 2031 different word forms in the corpus, totalling over 22.000 words in the monologues and over 47.000 in the dialogues, i.e. a grand total of about 70.000 words.

It is my distinct impression, shared by the project assistants who transcribe and annotate, that subjects were comfortable with the task and the experimental setting. They produced fluent speech for both monologues and dialogues and were not in any obvious way influenced by the non-naturalness of the circumstances.

4. Processing

Monologues and dialogues were transcribed orthographically in standard orthography, without punctuation, with capital letters for proper names only, with indication of empty and filled pauses, respectively, and with marks for articulatory hesitation.

The speech signals are segmented and annotated in Praat (http://www.fon.hum.uva.nl/praat/). The acoustic signal is segmented into prosodic phrases, words and syllables, always to the nearest zero-crossing in the waveform.



Figur 1: Praat screen print: waveform, spectrum and eight interval tiers. See further the text.

There are eight separate interval tiers for (1) the orthographical representation, (2) a detailed part-of-speech (POS) tagging, (3) a simplified POS-tagging, (4) an abstract phonological representation, (5) a broad phonetic transcription, (6) the pitch relation between each stressed and its first post-tonic syllable, (7) the phrasal intonation contour, and (8) an empty tier for comments. Fig. 1 is a screen print from one of the city-map monologues.

4.1. Annotation

The orthographical representation in tier 1 is supplemented with stress marks (commas directly before the vowel letter representing the vowel of the stressed syllable), intended for researchers who are interested only in the distribution of stress across the texts, regardless of the segmental pronunciation.

The POS-tagging in tiers 2 and 3 is automated. The tagger, developed by Peter Juel Henrichsen from the Department of Computer Linguistics at Copenhagen Business School, was trained on written language, not spontaneous speech (Henrichsen, 2002). At the outset there was no knowing how well the tagger would perform on non-scripted speech. However, although the tagger does make mistakes, they are not random. They are more or less confined to certain types, as revealed in the subsequent manual proof-reading process, and on the whole the tagger is efficient and reliable.

The phonological representation in tier 4 is fairly abstract where the segments are concerned, in accordance with the phonological analysis of Danish in Grønnum (2005), but stress marks are added to polysyllables, and stød is designated as well, although both stress and stød are to a very large extent predictable from the segmental and morphological structure and thus – strictly speaking – phonologically redundant. Adding stress and stød, however, will presumably facilitate certain search procedures at a later stage. (Stød is a special kind of creaky voice characterizing certain syllable types under certain morphological conditions. See, e.g., Grønnum and Basbøll, to appear.) The phonetic transcription in tier 5 is fairly broad.

The pitch relation in tier 6 between stressed and first post-tonic syllable is graded in seven steps: The post-tonic is much higher, higher, a little higher, equal, a little lower, lower, or much lower than the stressed syllable. The interval is specified to such relatively fine degree, because in its magnitude lies a correlate to perceived prominence (Grønnum, 1990; Jensen and Tøndering, 2005).

Phrasal intonationen in Danish, tier 7, is characterized by, firstly, the way the stressed syllables are pitch scaled

throughout the phrase, i.e. by their mutual relationship, and, secondly, presumably also by the way the phrase onsets and offsets, i.e. by the pitch of the very first and very last syllable, be it stressed or unstressed. The pitch of the stressed syllables and the syllables at the phrasal boundaries is represented on a coarse scale of high, mid or low. However, the means also exist to a finer gradation within a succession of stressed syllables in a given range (between high and mid, high and low, or mid and low). E.g., h_>_>_>_m designates a succession of five stressed syllables which descend gradually from high to mid.

Readers familiar with the ToBI convention for transcribing prosody (e.g., Silverman et al., 1992), should note that any similarity with our annotation is merely superficial. For the description of Danish intonation the phonological assumptions behind ToBI are inappropriate, and as a phonetic transcription system it is not sufficiently fine grained for our purpose (Grønnum 1985, 1986, 1995). For a general critique of ToBI, see Kohler (2005, 2006, to appear).

Note that, for reasons to do with time and resources, the pitch relation between successive prosodic phrases is not represented. Given the flexibility of Praat, it can easily be added to the grid if and when the need arises.

At the bottom is an empty tier for ad hoc comments.

The phonetic segmental and prosodic annotation in tiers 5-7 is performed independently and in parallel by two assistants. Disagreements between them are resolved in conferences with me. Subsequently, I go through and check the entire file. This procedure is repeated through every step: first the broad phonetic transcription, then the stress-and-pitch relation and finally the phrasal intonation.

5. Status

At the time of writing, the monologues have been annotated in their entirety, and a beta-version will shortly be available on the internet.

The dialogues have been orthographically transcribed, the sound files have been segmented at the syllable and word levels, the POS-tagging is complete, the phononological representation likewise, and we are well on the way with the segmental transcription. The entire corpus should be ready for publication by early 2007.

6. Acknowledgements

This project would not be possible, of course, without extensive help from many people, and not without external funding either. First and foremost, I am grateful to The Carlsberg Foundation for a grant of 1.25 million Danish kroner.

A number of individuals have each contributed invaluable assistance: Preben Dømler and Svend-Erik Lystlund assisted at the recordings. Gert Foget Hansen segmented a part of the monologues. Peter Juel Henrichsen supplied the POS-tagging and is also responsible for the search machine to accompany the corpus. Maja Dyrby and Line Burholt proof-read the POS-tags. John Tøndering transcribed orthographically all the monologues. He has written a number of immensely useful scripts for Praat, to locate mistakes, to move boundaries etc. He is also using the corpus for his own project and liberally shares his results with me. Nicolai Pharao supplied the 2031 word forms with an abstract phonological representation. Professor Hans Uszkoreit, Deutsches Forschungszentrum für Künstliche Intelligenz at Saarbrücken, has generously permitted me to borrow the licensing software from the TIGER Project (http://www.ims.uni-stuttgart.de/projekte/TIGER/). The major and most tedious work, however, is the responsibility of the transcribers, Cem Avus, Jeppe Beck, Andreas Geisler, Louise Astrid Johansson, Ruben Schachtenhaufen and Thit Wange Stærkær. Finally, without the twenty-seven speakers who gave liberally of their time and enthusiasm, none of this would have been possible.

7. References

Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S. and Weinert, R. (1991). The HCRC Map Task Corpus. Language and Speech 34, 4, pp. 351-366.

Brown, G., Anderson, A., Shillcock, R. and Yule, G. (1984). Teaching Talk. Cambridge: Cambridge University Press.

Christiansen, T.U. (2005). Byggestene i spontant talt dansk akustisk og perceptuelt. Unpublished project proposal.

Grønnum, N. (1985). Intonation and text in Standard Danish, Journal of the Acoustical Society of America 77, pp. 1205-1216.

Grønnum, N. (1986). Sentence intonation in textual context – supplementary data. Journal of the Acoustical Society of America 80, pp. 1040-1047.

Grønnum, N. (1990). Prosodic parameters in a variety of regional Danish standard languages, with a view towards Swedish and German. Phonetica 47, pp. 182-214.

Grønnum, N. (1995). Superposition and subordination in intonation – a nonlinear approach. In K. Elenius and P. Branderud (Eds.) Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm 1995, vol. II. Stockholm: KTH and Stockholm University, pp. 124-131.

Grønnum, N. (2005). Fonetik og Fonologi, 3. udg.,

København: Akademisk Forlag.

Grønnum, N. and Basbøll, H. (to appear). Danish Stød – Phonological and Cognitive Issues. In P.S. Beddor, M. Ohala, and M.-J. Solé Sabater (Eds.) Experimental approaches to Phonology. In honor of John J. Ohala. Oxford: Oxford University Press.

Henrichsen, P. Juel (2002). Sidste Års Aviser – Grammatisk opmærkning af et stort dansk aviskorpus. Lambda 27. København: Institut for Datalingvistik, Handelshøjskolen i København.

Jensen, C. and Tøndering, J. (2005). Choosing a Scale for Measuring Perceived Prominence. In Isabel Trancoso (Ed.) Proceedings of Interspeech 2005, September 4-8, Lisbon, Portugal, pp. 2385-2388.

Kohler, K.J. (2005). Timing and Communicative Functions of Pitch Contours. Phonetica 62, pp. 88-105.

Kohler, K.J. (2006). Paradigms in Experimental Prosodic Analysis – From Measurement to Function. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schließer (Eds.) Methods in Empirical Prosody Research. (= Language, context, and cognition, 3). Berlin, New York: de Gruyter.

Kohler, K.J. (to appear). Beyond Laboratory Phonology. In P.S. Beddor, M. Ohala, and M.-J. Solé Sabater (Eds.) Experimental approaches to Phonology. In honor of John J. Ohala. Oxford: Oxford University Press.

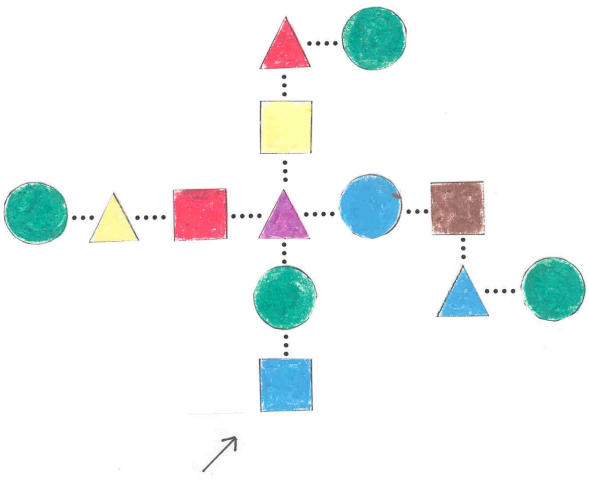
Silverman K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). ToBI: A standard for Labeling English Prosody. In Proceedings of the International Conference on Spoken Language Processing, pp. 867-870.

Swerts, M. (1994). Prosodic features of discourse units. Technische Universiteit Eindhoven.

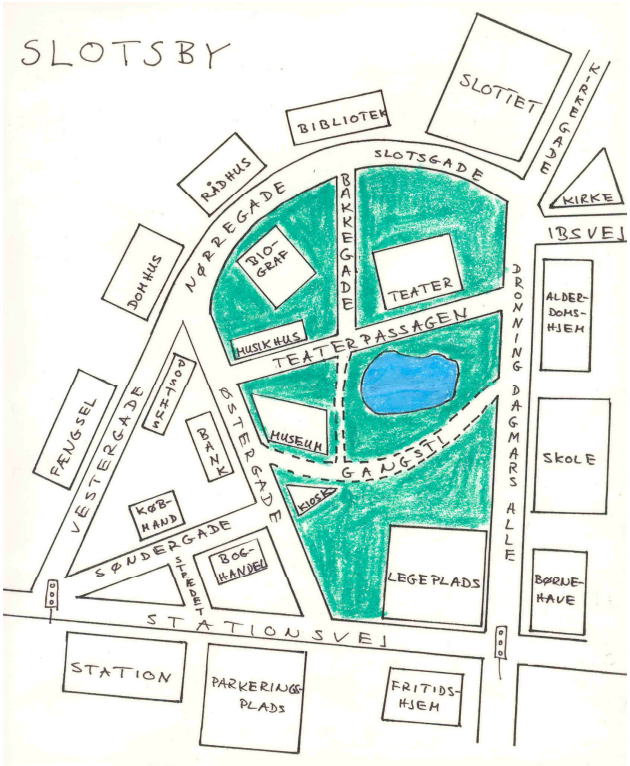
Swerts, M. and Collier, R. (1992). On the controlled elicitation of spontaneous speech. Speech Communication 121, pp. 463-468.

Terken, J.M.B. (1984). The distribution of pitch accents in instructions as a function of discourse structure. Language and Speech 27, pp. 269-289.

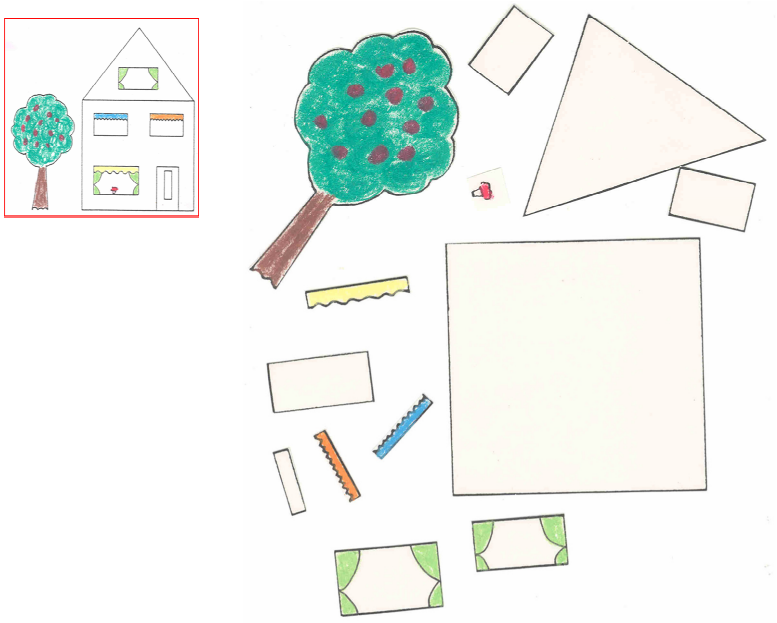
APPENDIX A



APPENDIX B



APPENDIX C



APPENDIX D