



SignSpeak Project

Scientific understanding and vision-based technological development for continuous sign language recognition and translation

Month 27 Evaluation Report

Major Deliverable D.7.2.M27

Release version: V1.0

Grant Agreement Number 231424

**Small or medium-scale focused research project (STREP)
FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics**

Project start date: 1 April 2009

Project duration: 36 months

Dissemination Level		
PU	Public (can be made available outside of SignSpeak Consortium without restrictions)	X
RE	Restricted to SignSpeak Programme participants and a specified group outside of SignSpeak consortium	
IN	SignSpeak Internal (only available to (all) SignSpeak programme participants)	
LI	SignSpeak Limited (only available to a specified subset of SignSpeak programme participants)	
Distribution list (only for RE or LI documents)		

0 General Information

0.1 Document

Title	Month 27 Evaluation Report
Type	Major Deliverable
Ref	D.7.2.M27
Target version	V1.0
Current issue	V0.3
Status	Draft
File	D.7.2.M27.EvaluationReport.tex
Author(s)	Jens Forster, Yannick Gweth and Hermann Ney / RWTH Justus Piater, Du Wei, and Thomas Hoyoux / UIBK Gregorio Martínez, Jaume Vergés-Llahí, and Juan D. García-Arteaga / CRIC
Reviewer(s)	Gregorio Martinez / CRIC, Jens Forster / RWTH
Approver(s)	Gregorio Martinez / CRIC
Approval date	
Release date	04/08/2011

0.2 History

Date	Version	Comment
04/08/2011	V0.3	handed out version for quality control
02/08/2011	V0.2	incorporated material from partners
15/07/2011	V0.1	first description of D.7.2 report

0.3 Document scope and structure

The tasks in WP7 are intended to evaluate the deliverables generated within the technical work packages (WP3, WP4 and WP5), with the aim of providing a constant monitoring of progress and obtaining feedback for the next developments. The document consists of a first part describing the objectives of the project in the present project, followed by the evaluation of the different elements composing each WP, namely, the multi-modal visual analysis and the sign language recognition and translation tasks.

Authors	Group
Forster, Jens	RWTH
Gweth, Yannick	RWTH
Koller, Oscar	RWTH
Ney, Hermann	RWTH
Schmidt, Christoph	RWTH
Zelle, Uwe	RWTH
Piater, Justus	UIBK
Wei, Du	UIBK
Hoyoux, Thomas	UIBK
Martinez, Gregorio	CRIC
Vergés-Llahí, Jaume	CRIC
García-Arteaga, Juan D.	CRIC

0.4 Content

0 General Information	2
0.1 Document	2

0.2	History	2
0.3	Document scope and structure	2
0.4	Content	2
1	Project Objectives for the Period	4
2	Technical Accomplishment	4
2.1	Multimodal Visual Analysis (Task 7.1)	5
2.1.1	Workpackage objectives and starting point at the beginning of the period	5
2.1.2	Progress towards objectives	5
2.1.3	Evaluation of the Multimodal Visual Analysis	5
2.2	Sign Language Recognition (Task 7.2)	7
2.2.1	Workpackage objectives and starting point at the beginning of the period	7
2.2.2	Progress towards objectives	8
2.2.3	Evaluation of Sign Language Recognition	8
2.3	Sign Language Translation (Task 7.3)	13
2.3.1	Workpackage objectives and starting point at the beginning of the period	13
2.3.2	Progress towards objectives	13
2.3.2.1	Task 5.1 Translation from gloss-based corpora	13
2.3.2.2	Task 5.2 Handling translation of parallel information channels	16
2.3.2.3	Task 5.3 Handling of spatial reference points	16
2.3.2.4	Task 5.4: Handling of temporal signs and incorporations	17
2.3.2.5	Task 5.5: Comparing parallel input channel approaches for multi-modal translation	17
2.3.3	Clearly significant results	18
3	Objectives for the next Evaluation	19
4	References	19

Table 1: Expected corpus annotation progress of the RWTH-PHOENIX and Corpus-NGT corpora in comparison to the limited domain speech (Verbmobil II) and translation (IWSLT) corpora.

	BOSTON-104	Phoenix		Corpus-NGT		Vermobil II	IWSLT
year	2007	2011	2012	2009	2011	2000	2006
recordings	201	check	400	116	300	-	-
running words	0.8k	20k	50k	30k	80k	700k	200k
vocabulary size	0.1k	0.8k	< 2.5k	3k	< 5k	10k	10k
T/T ratio	8	20	> 20	10	> 20	70	20
Performance	11% WER [3]	-	-	-	-	15% WER [7]	40% TER [11]

1 Project Objectives for the Period

The tasks in WP7 are intended to evaluate the deliverables generated within the technical work packages (WP3, WP4 and WP5), with the aim of providing a constant monitoring of progress and obtaining feedback for the next developments.

For the second period, prototypes have been improved for the multi-modal visual analysis (D.3.2), the sign language recognition (D.4.2) and sign language translation (D.5.2). This deliverable D7.2 gathers the evaluation of these three prototypes.

Major achievements in reaching scientific and technological project objectives were:

- Evaluation results of the extended prototypes developed during the second period of the SignSpeak project are presented in this report.
- Robust tracking algorithms are required as the signing hand frequently moves in front of the face, may temporarily disappear, or cross the other hand. Features based on robust tracking are one of the key elements for marker-less sign language recognition.
- Only few studies consider the recognition of continuous sign language, and usually special devices such as colored gloves or blue-boxing environments are used to accurately track the regions-of-interest in sign language processing.
- Ground-truth labels for head-shakes have been annotated for several sentences of Corpus-NGT. The annotation of ground-truth labels for RWTH-PHOENIX-v2.0 and SIGNUM databases has been finished. Tracking error rates below 12% for the dominant hand and below 1% have been achieved for the RWTH-PHOENIX-v2.0, SIGNUM and RWTH-BOSTON-104 databases. Furthermore, signer-dependent recognition result in the range of 20% word error rate (WER) have been achieved on the SIGNUM database. Finally, translation results for NGT to Dutch and DGS to German are presented.

2 Technical Accomplishment

The present section will establish the kind of data employed throughout the document in order to evaluate and validate the performance of the different WPs.

In order to build a Sign-Language-to-Spoken-Language translator, reasonably sized corpora have to be created for statistically-based data-driven approaches. For a limited domain speech recognition task (Verbmobil II) as e.g. presented in [7], systems with a vocabulary size of up to 10k words should be trained with at least 700k words to obtain a reasonable performance, i.e. about 70 observations per vocabulary entry. Similar values should be obtained for a limited domain translation task (IWSLT) as e.g. presented in [11].

Similar corpora statistics can be observed for other ASR or MT tasks. The requirements for a sign language corpus suitable for recognition and translation can therefore be summarized as follows:

- annotations for a limited domain (e.g. broadcast news, etc.)
- for a vocabulary size smaller than 4k words, each word should be observed at least 20 times
- the singleton ratio should ideally stay below 40%

Existing corpora should be extended to achieve a good performance w.r.t. recognition and translation [6]. During the SignSpeak project, the existing RWTH-PHOENIX corpus [14] and Corpus-NGT [1] will be extended to meet these demands (c.f. Table 1).

2.1 Multimodal Visual Analysis (Task 7.1)

The objective of this task is to provide a quantitative and partly qualitative evaluation of the individual components for visual analysis developed under WP3 by this time.

2.1.1 Workpackage objectives and starting point at the beginning of the period

2.1.2 Progress towards objectives

During the second period RWTH has focussed on two major activities in WP3. The first activity has been the extraction of features suitable for hand/head tracking and sign language recognition. Those features are local features such as SIFT and SURF, as well as geometric features. Furthermore, RWTH's second major activity has been in the area of hand and head tracking and software development for tracking.

UIBK has pursued similar objectives, namely the development of tools for hand and head tracking, and extraction of features that are suitable for sign language recognition. However UIBK has exploited different methods for tracking (Active Appearance Models, Linear Programming, ...) and has tried to achieve feature extraction of higher semantical level (eyebrow raising, ...).

2.1.3 Evaluation of the Multimodal Visual Analysis

In order to evaluate the performance of head and hand tracking algorithms, data with ground-truth annotations is required as well as an evaluation measure.

For an image sequence $X_1^T = X_1, \dots, X_T$ and corresponding annotated object positions $u_1^T = u_1, \dots, u_T$, the tracking error rate (TER) of tracked positions \hat{u}_1^T is defined as the relative number of frames where the Euclidean distance between the tracked and the annotated position is larger than or equal to a TER tolerance τ :

$$\text{TER} = \frac{1}{T} \sum_{t=1}^T \delta_{\tau}(u_t, \hat{u}_t) \quad \text{with} \quad \delta_{\tau}(u, v) := \begin{cases} 0 & \|u - v\| < \tau \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

In contrast to the dpt approaches, a model-based face detection approach is used for comparison at RWTH, where the faces have been automatically detected using the OpenCV implementation of the Viola & Jones [17] face detector, independently in each frame (named Viola&Jones Detector (VJD) in this section). The search-space is constrained in an improved tracking version by successful predecesing head detections (i.e. Kalman filter-like, named Viola&Jones Tracker (VJT) in this section) in order to avoid confusions with noisy head detections, e.g. in the background. It uses the predecesing head detection in case of no-detection. As the cascades have been trained on different data, the detection approach is again model-based but person-independent. The coordinates of the detection rectangle's center is used for tracking evaluation.

The POICAAM-based face tracker from UIBK is evaluated w.r.t to the TER by comparing the nose tip landmark positions obtained by tracking in each frame, to the nose tip ground truth previously annotated by RWTH. It should be noted that the nose tip landmark is only one of the many points (typically 40) tracked on the face within the POICAAM framework; in the absence of more detailed ground truth for SignSpeak datasets, comparing the nose tip positions just gives a rough quantitative evaluation of the tracking performance. In order to asses the quality of feature extraction at UIBK - the mouth aperture, eyes aperture and level of eyebrows raising are extracted using 1NN shape-based classifiers (see the technical report D.3.4) - some frames (200 images) of the RWTH-PHOENIX-v2.0 dataset have been annotated for training, and eight full videos (1722 images) of the same dataset have been annotated for testing. Those annotations consist of classes for the features of interest: mouth/eyes can be in one of four classes - shut, almost shut, open, wide-open (0,1,2,3) - eyebrows can be in one of three classes - frown, relaxed, raised (0,1,2). Confusion matrices and related measures have been computed for two scenarios: (a) taking all classes into account, and (b) merging classes to have binary classifiers: mouth/eyes can be open or shut, eyebrows can be raised or not raised; see Table 6 (note that only person specific models were used for this evaluation). From the POICAAM-based tracking results, UIBK has also extracted the head pose in 3D space as a feature for sign language recognition. This could however not be evaluated because of the absence of appropriate ground truth.

Corpus-NGT. RWTH annotated about 8k frames. Hand and head tracking results on the Corpus-NGT tracking snapshot database comparing model-free and -dependent approaches are given in Table 2. Main problems encountered on this database are fast and abrupt hand and head movements, as well as changing and uncontrolled environment conditions. Spatial pruning and data preprocessing steps found to benefit tracking performance for the RWTH-PHOENIX-v2.0 database are currently being investigated for Corpus-NGT.

RWTH-PHOENIX-v2.0. RWTH annotated the spatial positions of the hands and the head of seven different signers in 39,691 images of the RWTH-PHOENIX-v2.0 database and distributed this data to the partners. For

Table 2: Hand and head tracking results on the Corpus-NGT tracking snapshot database

Tracking	Model	Pers. dep.	# Frames	Setup	TER [%]			
					$\tau=5$	$\tau=10$	$\tau=15$	$\tau=20$
Dominant Hand	no	no	7891	DPT (RWTH)	86.80	63.24	47.26	37.99
	yes	no	7891	ASCS (CRIC)	92.11	75.08	56.20	41.16
Non-Dominant Hand	yes	no	7891	ASCS (CRIC)	89.82	69.72	53.28	42.62
Head	yes	no	7891	DPT+PCA (RWTH)	98.18	92.13	75.82	59.43
	yes	no	7891	DPT+VJ (RWTH)	69.56	30.91	16.72	12.17
	yes	no	7891	VJD (RWTH)	78.13	62.07	59.59	58.52
	yes	no	7891	VJT (RWTH)	56.92	26.04	17.55	15.81
	yes	no	7891	ASCS (CRIC)	93.05	82.00	65.40	45.36

easy access, the annotation statistics are subsumed in Table 3. Hand and head tracking results on the RWTH-PHOENIX-v2.0 tracking snapshot database comparing model-free and -dependent approaches are given in Table 5.

Table 3: Tracking annotation statistics RWTH-PHOENIX-v2.0 database

number of frames annotated	number of sentences	number of τ different signers
39,691	266	7

Using the dynamic programming tracking framework [2] and motion based scoring functions, RWTH performed tracking experiments on the RWTH-PHOENIX-v2.0 database. All experiments have been performed at half of the original video resolution, i.e. 105×130 , using a sliding window of size 25×35 which corresponds to the average minimum rectangle enclosing the complete dominant hand of a signer.

Besides the image resolution and sliding window size, the experiments using gray respectively color images use the same scoring functions and parameter setting as used in the best tracking setup of the RWTH-BOSTON-104 database. Baseline tracking results of 63.98% TER for gray images and 57.48% TER for color images at $\tau = 20$ obtained on the RWTH-PHOENIX-v2.0 database shown in Table 4, which use the same tracking configuration as e.g. the RWTH-BOSTON-104 database, show that additional preprocessing has to be done in order to achieve reasonable tracking error rates.

An analysis of the tracking errors indicated that the majority of tracking errors was due to tracking the non-dominant instead of the dominant hand and switching of hands in the case of hand interchange. In order to reduce these kind of errors, RWTH tested two types of spatial pruning.

First, hard spatial pruning forces the algorithms to avoid areas left of the body axis in tracking of the dominant hand. Hard spatial pruning leads to an absolute improvement of $\approx 7\%$ for gray and color images. Additional brightness normalization by power law gives between 2% and 1% improvement. Because results for color images are consistently $\approx 7\%$ better than results for gray images, RWTH focussed on exploiting color and skin cues for tracking. RWTH exploited skin color information to suppress non skin-colored image parts for the calculation of motion scores. Doing so, an absolute improvement of 35% was achieved. Relaxing the hard constraint of hard spatial pruning, called smart pruning, allows the tracking algorithm to consider image areas remote from the body axis albeit at an increasing penalty. In the described experiments the penalty is governed by a linear function close to the body axis and by an exponential function on the opposing side of the body axis, e.g. on the left side of the body axis for the right hand. Experiments using smart spatial pruning improve over hard spatial pruning by 0.5%. In combination with the suppression of non skin-colored image areas the gain is about 1.5%. Detailed preliminary tracking results by RWTH are subsumed in Table 4.

The best preliminary tracking results for hand and head tracking by CRIC, RWTH, and UIBK are subsumed in Table 5.

The POICAAM-based face tracker gives the best performance w.r.t the TER. In particular on the RWTH-PHOENIX-v2.0 dataset this method outperforms the other ones, both in the person dependent and the person independent setup (i.e. using one different model for each signer vs. one model for all signers of the dataset). The person independent setup even gives better results than the person dependent setup; this last point contradicts the theory on AAMs, and this evaluation artifact is mainly due to the fact that the tracking evaluation conducted here considers only the position of the nose tip.

SIGNUM. RWTH annotated about 52k frames of the SIGNUM database with spatial positions of the hands and the face. The relevant statistics of this tracking subset of the SIGNUM database are presented in Table 7.

Table 4: Preliminary DPT tracking results for RWTH-PHOENIX-v2.0 by RWTH

Setup	TER [%]			
	$\tau=5$	$\tau=10$	$\tau=15$	$\tau=20$
Gray	96.48	87.18	75.17	63.98
+ Hard Spatial Pruning	95.76	84.35	70.35	57.01
+ Brightness Normalization	95.52	83.52	68.86	55.35
Color	94.84	82.23	68.35	57.48
+ Hard Spatial Pruning	94.16	79.56	62.94	49.96
+ Brightness Normalization	93.76	79.10	62.42	48.70
+ Non-Skin Suppression	84.32	52.89	27.16	13.69
+ Smart Spatial Pruning	94.18	80.19	63.68	49.58
+ Non-Skin Suppression	84.17	52.89	26.74	12.50
+ Face Suppression	82.92	51.13	25.14	11.68

Table 5: Hand and head tracking results for the RWTH-PHOENIX-v2.0 tracking snapshot database

Tracking	Model	Pers. dep.	# Frames	Setup	TER [%]			
					$\tau=5$	$\tau=10$	$\tau=15$	$\tau=20$
Dominant Hand	no	no	39691	DPT (RWTH)	82.92	51.13	25.14	11.68
	yes	no	39691	ASCS (CRIC)	87.97	71.84	57.04	46.38
Non-Dominant Hand	yes	no	39691	ASCS (CRIC)	79.07	57.32	43.75	36.18
Head	yes	no	39691	DPT+VJ (brightness-sum)	74.13	19.55	3.96	1.34
	yes	no	39691	DPT+VJ (skinprob-sum)	74.95	19.63	3.09	0.44
	yes	no	39691	VJD (RWTH)	77.50	20.64	13.19	13.02
	yes	no	39691	VJT (RWTH)	66.34	17.21	7.45	4.23
	yes	no	39691	ASCS (CRIC)	77.42	36.29	10.89	3.13
	yes	no	39691	POICAAM (UIBK)	6.44	0.69	0.27	0.15
	yes	yes	39691	POICAAM (UIBK)	7.48	0.88	0.24	0.08

Actually the position of head, right and left hand of 3 different signers have been annotated for the evaluation of the signer independent setup. Preliminary tracking results on the SIGNUM database are shown in Table 8.

On the full labeled dataset, RWTH achieves 10.7% TER for a 15×15 search window, where frames in which the hands are not visible, are not considered.

RWTH-BOSTON-104. RWTH annotated about 15k frames. Hand and head tracking results on the RWTH-BOSTON-104 database comparing model-free and -dependent approaches are given in Table 9. It can be observed that the DPT approach outperforms all other evaluated approaches, both for head and hand tracking. The DPT+ZOW approach can in comparison to the baseline DPT approach further improve the tracking accuracy, especially for smaller τ tolerance values. The Viola&Jones+Kalman tracking approach by RWTH significantly outperforms the baseline DPT+PCA Eigenface tracking approach, and provides a very good tracking accuracy even for small τ tolerance values.

Again on the RWTH-BOSTON-104 database, the POICAAM-based face tracker outperforms the other methods when evaluated w.r.t the TER. Here, only a person independent setup has been evaluated.

2.2 Sign Language Recognition (Task 7.2)

The objective of this task is to provide a quantitative and partly qualitative evaluation of the individual components for visual analysis developed in WP4 by this time.

2.2.1 Workpackage objectives and starting point at the beginning of the period

Based on the Corpus-NGT Isolated Snapshot V0.1 and RWTH-BOSTON-50 databases, RWTH has defined training and test sets for signer dependent and signer independent recognition of isolated signs to be used in the M27 evaluation. Due to visual complexity of the Corpus-NGT data (see WP3 and WP4 for details), only the SIGNUM, RWTH-BOSTON-104 and RWTH-PHOENIX-v2.0 corpora will be used at the current stage of the project for the evaluation of continuous sign language recognition. RWTH has defined training, development and testing datasets for the RWTH-PHOENIX-v2.0 and the SIGNUM databases. While the SIGNUM database has been used for internal evaluation of the developed recognition system only, RWTH distributed the relevant training, development and test sets of the RWTH-PHOENIX-v2.0 database to CRIC and UIBK.

Main objectives for the second period were

Table 6: Feature classification accuracy (%) on the RWTH-PHOENIX-v2.0 database

	Scenario A (3-4 classes)	Scenario B (2 classes)
Mouth aperture	56.39	79.50
Eyes aperture	58.25	80.55
Eyebrows raising level	72.13	83.22

Table 7: 286 segments annotated with head and hand ground-truth positions for tracking

SIGNUM	Annotated Frames
Training	38012
Development	6686
Test	6750
All	51448

- to establish database snapshots for
 - RWTH-PHOENIX-v2.0
 - SIGNUM multi signer
- to establish baseline results for
 - continuous sign language recognition

2.2.2 Progress towards objectives

A major activity in this period was the development of the extended sign language recognition prototype. This prototype deliverable D.4.2 constitutes the second implementation of the sign language recognition module in the SignSpeak data pipeline as described in Section B.1.3.1 and WP6 Task 6.1. of the Technical Annex Document. This extended prototype is currently being used for several corpora and sign languages:

- RWTH-BOSTON-50 and RWTH-BOSTON-104 corpora (American Sign Language (ASL))
- Corpus-NGT corpus (Nederlandse Gebaren Taal (NGT))
- SIGNUM and RWTH-PHOENIX-v2.0 corpora (Deutsche Gebärdensprache (DGS))

Experiments using whole-word modeling have been performed with the extended prototype on the RWTH-BOSTON-50 (ASL) and Corpus-NGT (NGT) databases for isolated sign language recognition to establish the work flow between the different work packages. Experiments for continuous sign language recognition concerning **pronunciation, language modelling adaptation and the usage of speaker adaptation techniques** have been analyzed for ASL on the RWTH-BOSTON-104 database, for DGS on the RWTH-PHOENIX-v2.0 and SIGNUM database.

During the second period RWTH has focussed on two major activities in WP4. The first activity has been the preparation of **external feature integration** within **multiple corpora setups for sign language recognition**. The second major activity has been the extension of the sign language recognition prototype for **continuous sign language recognition**, especially adaptation of discriminative training and feature extraction approaches.

The D.4.2 report gives an overview of results obtained with the extended prototype system for ASL and DGS. **The extended prototype can now be used for a variety of video types and languages**. The full work-flow pipeline of WP3, WP4, and WP5 prototypes has been published in [5] at the LREC 2010 conference, with a more detailed description of the corpora to be used in [4]. To measure the quality of externally generated features a tracking evaluation has been performed (c.f. Task 7.1).

Furthermore, RWTH has carried out an extensive evaluation of the sign language recognition system and features for the RWTH-PHOENIX-v2.0, SIGNUM, and RWTH-BOSTON-104 databases.

2.2.3 Evaluation of Sign Language Recognition

All developed methods necessary for isolated and continuous sign language recognition are measured, similarly to speech recognition, in terms of WER. For isolated sign language recognition the WER is simply the error rate,

Table 8: Hand and head tracking results on the SIGNUM tracking subset

Tracking	Model	Pers. dep.	# Frames	Setup	TER [%]			
					$\tau=5$	$\tau=10$	$\tau=15$	$\tau=20$
Dominant Hand	no	no	(all) 51448	DPT (RWTH)	85.5	53.8	24.9	10.7
Head	yes	no	(all) 51448	DPT+PCA (RWTH)	94.06	61.48	25.36	9.02
	yes	no	(all) 51448	DPT+VJ (RWTH)	56.47	13.47	2.18	1.08
	yes	no	(all) 51448	VJD (RWTH)	80.57	32.56	10.80	6.90
	yes	no	(all) 51448	VJT (RWTH)	79.84	29.56	5.93	1.51

Table 9: Hand and head tracking results on the RWTH-BOSTON-104 tracking subset

Tracking	Model	Pers. dep.	# Frames	Setup	TER [%]			
					$\tau=5$	$\tau=10$	$\tau=15$	$\tau=20$
Dominant Hand	no	no	12909	DPT (RWTH)	73.59	42.29	18.79	8.37
	no	no	12909	DPT+ZOW (RWTH)	75.44	36.36	14.51	8.06
	no	no	2603	DPT (RWTH)	74.79	44.33	20.43	8.83
	yes	yes	2603	Robust PCA (ULG)	89.86	77.41	64.50	47.48
Non-Dominant Hand	yes	yes	842	Robust PCA (ULG)	80.19	57.78	39.39	24.06
Head	yes	no	15732	DPT+PCA (RWTH)	26.77	17.32	12.70	10.86
	yes	no	15732	DPT+VJ (RWTH)	10.06	0.40	0.02	0.00
	yes	no	15732	VJD (RWTH)	9.75	1.23	1.09	1.07
	yes	no	15732	VJT (RWTH)	10.04	0.81	0.73	0.68
	yes	yes	15732	AAM (ULG)	10.17	6.85	6.82	6.81
	yes	no	15732	AAM (ULG)	10.92	7.92	7.88	7.76
	yes	no	15732	POICAAM (UIBK)	3.54	0.12	0.08	0.08

but for continuous sign language recognition the WER is composed of errors that are due to deletion, insertion, or substitution of words:

$$\text{WER} = \frac{\#deletions + \#insertions + \#substitutions}{\#observations} \quad (2)$$

RWTH-BOSTON-104 As our corpora are annotated in glosses, i.e. whole-word transcriptions, the system is based on whole-word models. Each word model for the RWTH-BOSTON-104 database consists of one to three pseudo-phonemes modeling the average word length seen in training. Our RWTH-BOSTON-104 lexicon defines 247 pseudo-phonemes for 104 words. Each pseudo-phoneme is modeled by a 3-state left-to-right hidden Markov model (HMM) with three separate Gaussian mixtures (GMM) and a globally pooled covariance matrix. Preliminary recognition results comparing different tracking approaches, features, and training approaches are presented in Table 10. It can be observed that modified-MMI (M-MMI) based discriminative training and multi-layer perceptronX (MLP) based features do not yet behave as in other domains such as automatic speech recognition (ASR), where larger amounts of training data are typically used. From the difference between the unsupervised tracking DPT hand-patch results and the supervised generated hand-patch results (33.71% WER vs. 30.34% WER) we can conclude that the quality of the considered DPT tracker is of sufficient quality. **The PCA and LDA reduced AAM features [12] provided by UIBK demonstrate the ongoing integration work of WP3 features into the RWTH recognition framework.** Currently all coordinates are used as initial features which are reduced by PCA or LDA, more interesting will be a manual feature selection (c.f. WP2 expert knowledge) in the future. Finally, a combination of all existing features is still missing and part of our RWTH's ongoing research efforts.

SIGNUM Signer dependent recognition experiments have been carried out on the SIGNUM database using the extended recognition prototype. Table 11 shows the corpus statistics for this setup. Three successive feature vectors are concatenated to a large input feature vector. RWTH applies Principal Component Analysis (PCA) on this feature vector. PCA is used to reduce the dimension of the large feature vector keeping only the most discriminative elements. This windowing technique improves the WER from 38.5% to 33.8% as shown in Table 12. A larger window size shows no improvement of the error rate so far.

Furthermore, geometric features describing the hand shape and configuration of the dominant-hand have been extracted and combined with appearance based features using PCA based feature combination method. The WER of the prototype using the geometric features are reported in Table 13.

A log-linear model combination technique[18] has been used to take advantage of the mutual information provided by the appearance-based features as well as the dominant hand and the facial expression features.

Table 10: Hand and head features on the RWTH-BOSTON-104 dataset (MFDI results achieved in [13])

Tracker	Features	del	ins	sub	errors	WER [%]
-	Frame (1024)	39	10	20	69	38.76
-	PCA-Frame (110)	20	9	20	49	27.53
DPT (RWTH)	Dom. Hand-Patch (1024)	27	8	31	66	37.08
	+ PCA (30)	17	13	30	60	33.71
Ground-truth	+ PCA (30)	9	12	33	54	30.34
DPT (RWTH)	PCA-Frame (110)					
	+ hand-position u_t (112)	31	1	13	45	25.28
	+ hand-motion m_t with $\Delta \in 1, 2$ (114)	28	2	16	46	25.84
	+ hand-trajectory eigenvalues (112)	20	4	19	43	24.16
DPT+PCA (RWTH)	MFDI (1024)	-	-	-	-	56.2
	+ PCA (110)	-	-	-	-	54.0
AAM (UIBK)	AAM Face Coordinates (80)	32	9	42	83	46.63
	+ PCA (30)	26	12	36	74	41.57
	+ LDA (30)	42	11	38	91	51.12

Table 11: SIGNUM database signer dependent setup statistics

	Trainset	Testset
# signers	1	1
# overlap	-	all
# shows	1809	531
# sentences	1809	531
# frames	416620	-
# running words	11109	2805
vocabulary size	569	400
# sub-units	455	387
# OOV [%]	-	0.1
avg. sentence length [glosses]	6.1	5.3
avg. gloss length [frames]	23.2	-
TTR	16.2	-
perplexity (3-gram)	17.8	72.2

Table 13 shows the results reached by a log-linear combination of the best independent models during the recognition process. In additional experiments, facial features as given by their coordinates have been extracted and trained separately. Log-linear combination using this third model in addition to the two previous ones improves the recognition to 20.3%. Further WP3 feature and WP4 model combinations are expected to achieve improvements in the future.

In addition to the signer-dependent recognition experiments on the SIGNUM database RWTH carried out multi-signer experiments using all 25 different signers of the SIGNUM database. The statistics of the multi-signer SIGNUM setup are subsumed in Table 14.

Preliminary results on the multi-signer setup of the SIGNUM database using the same baseline setup as in the signer dependent setup yields an WER of 47.1. Further experiments using the multi-signer setup of the SIGNUM database is ongoing work at RWTH.

Corpus-NGT Isolated Snapshot V0.1 and RWTH-PHOENIX-v2.0 Preliminary recognition experiments for multi signer sign language recognition for isolated signs (Corpus-NGT Isolated Snapshot V0.1 corpus) and continuous sign language (RWTH-PHOENIX-v2.0) carried out by RWTH show so far poor recognition performance. Isolated Snapshot V0.1 suffers from challenging video conditions, poor hand tracking performance and training alignment issues. Similar alignment effects are observed for the RWTH-PHOENIX-v2.0 corpus and detailed in the remainder of this section.

Table 15 shows the statistics for train, development, and test sets of the first recognition snapshot of the

Table 12: Baseline recognition experiments on the SIGNUM dataset

Features	Feature vector size	del	ins	sub	errors	WER [%]
Frame	1024	896	66	1132	2094	74.7
+ PCA (200)	200	311	133	636	1080	38.5
+ win 3 + PCA (200)	200	184	148	616	948	33.8

Table 13: PCA and log-linear combination of appearance-based, geometric and facial features on the SIGNUM dataset

Tracker	Features	del	ins	sub	errors	WER [%]
-	Frame + win3 + PCA (200) (model-1)	184	148	616	948	33.8
DPT (RWTH)	Handpatches					
	+ win3 + PCA (200) (model-2)	156	132	371	659	23.5
	+ win3 + PCA (200) + geometric (model-3)	130	184	512	826	29.4
DPT (RWTH)	Frame					
	+ PCA (200) + geometric (34)	201	113	557	871	31.0
	+ win3 + PCA (200) + geometric (34)	121	139	531	801	28.5
VJT (RWTH)	STASM facial coordinates					
	+ PCA (100) (model-4)	1036	105	1304	2445	87.2
	log-linear model combination					
	model-1 + model-2 + model-4	146	113	420	679	24.2
	model-1 + model-2 + model-3	176	64	329	569	20.3

Table 14: SIGNUM database multi signer setup statistics

	Trainset	Testset
# signers	25	25
# overlap	-	all
# shows	15075	4425
# sentences	15075	4425
# frames	3.6 10 ⁶	-
# running words	92575	23375
vocabulary size	455	-
# OOV [%]	-	0.1
avg. sentence length [glosses]	6.1	5.3
avg. gloss length [frames]	23.2	-
TTR	203.4	-
perplexity (3-gram)	17.8	72.2

RWTH-PHOENIX-v2.0 database. All seven signers appear in the training, development, and test set.

Table 15: RWTH-PHOENIX-v2.0 continuous recognition snapshot version 0.1

	Train	Development	Test
# sentences	1,054	175	175
# running glosses	11,665	1,900	1,801
vocab. size	521	—	—
# singletons	219	—	—
OOV%	—	0.84	1.01
avg. sentence length	11.06	10.85	10.29
avg. gloss length [frames]	13.2	—	—

RWTH has taken action to ensure that the class and signer distributions between training, development, and test sets are approximately the same. The average type-token-ratio (TTR) of the training set is 22.3 albeit being biased by 219 glosses which are seen only once during training and a long-tail in the frequency distribution i.e. the 20 most frequent signs explain more than 50% of all data. Nevertheless, RWTH-PHOENIX-v2.0 shows a trigram perplexity of 55.9 on the test set.

As mentioned before, the recognition system so far shows poor performance on the RWTH-PHOENIX-v2.0 dataset. RWTH carried out a qualitative evaluation on the alignments between the extracted features and the states of the left-to-right HMM models.

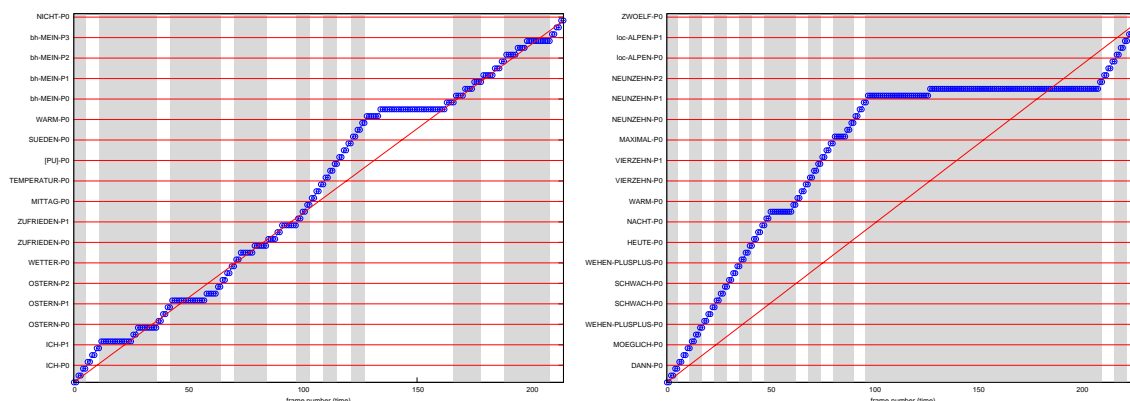


Figure 1: Visualization of Time Alignment for 2 Train Sentences in RWTH-PHOENIX-v2.0 database using Gaussian Density models

Figure 1 shows the temporal HMM state emission alignment for two training sentences of the RWTH-PHOENIX-v2.0 database. The x-axis is labeled by the frame number of the video segment, which correspond to the relative time with regard to the segment start, and the y-axis is labeled by the state number of the corresponding pseudo-phoneme class. The alternating white and grey background illustrate the temporal pseudo-phoneme boundaries, the horizontal red lines are optical reference lines indicating the last aligned state of a pseudo-phoneme, and the red diagonal is a optical reference line for the theoretically optimal training alignment. Each blue circle indicates an aligned state of the corresponding pseudo-phoneme. The system uses six HMM states and three tied Gaussian single density emission models for each pseudo-phoneme, e.g. states 0 and 1 share the same distribution. RWTH tested several appearance and tracking based featureres resulting in similar alignments as depicted in Figure 1.

While the left alignment in Figure 1 is pretty close to the theoretically optimal state alignment apart from about 40 frames aligned to the first pseudo-phoneme of class "bh-mein", the alignment shown on the right side of Figure 1 is clearly not desirable in statistical recognition system. The majority of pseudo-phonemes has been aligned only to the minimum number of frames to visit all states in the alignment on the right hand side while the pseudo-phoneme "NEUNZEHN-P2" is aligned to more than 100 consecutive frames. Unfortunately, the majority of alignments for the RWTH-PHOENIX-v2.0 database shows strong deviations from the theoretically optimal alignment path leading to visual models that do not explain the training data well. Hence, it is not surprising that recognition results for the RWTH-PHOENIX-v2.0 database do currently not allow for sensible interpretation. As a first step to improve on the shown alignments RWTH annotated isolated signs to learn the average length of

each sign from ground truth data. So far the integration of this length modelling has not improved the alignment situation.

RWTH is currently evaluating features from WP3 with regard to alignment quality and resulting recognition performance. A major challenge in the RWTH-PHOENIX-v2.0 database is the temporal resolution of the video sequence.

2.3 Sign Language Translation (Task 7.3)

2.3.1 Workpackage objectives and starting point at the beginning of the period

The primary objective of this workpackage is to develop sign language translation technologies which perform an automatic machine translation of multimodal input from recognized signs transcribed in gloss notation into a spoken language. The translation system should deal with the parallel and multimodal nature of sign languages.

2.3.2 Progress towards objectives

The main activity in this period was the development of an advanced prototype for sign language translation. Moreover, first steps were taken to tackle multimodal input. We performed the following activities:

- Databases:
 - Further annotation of the RWTH-PHOENIX-v2.1 to obtain more training data for the translation experiments
 - Discussion of translation issues in Corpus-NGT to improve translations
- Software:
 - Setting up of automatic translation pipelines for both the phrase-based and the hierarchical system
 - Implementation of a technique similar to cross-validation to stabilize optimization on small corpora
 - Incorporation of lexical knowledge by using Morphisto¹ results for the training of word alignments
 - Parsing of eaf files and preprocessing of the corpora to handle multi-tier input in machine translation
- Experiments:
 - Experiments on RWTH-PHOENIX-v2.0 in both translation directions
 - Comparative experiments on RWTH-PHOENIX-v2.1 from glosses to Spoken German
 - Experiments on Corpus-NGT in both translation directions

We performed experiments on the RWTH-PHOENIX-v2.0 and on Corpus-NGT, using both our in-house phrase-based decoder as well as the hierarchical phrase-based decoder Jane, which has become an open-source software [16].

2.3.2.1 Task 5.1 Translation from gloss-based corpora

RWTH performed experiments on the RWTH-PHOENIX-v2.0 translation corpus, using both its in-house phrase-based decoder and its recently published open-source hierarchical decoder Jane. At the end of the section, new results on the latest snapshot RWTH-PHOENIX-v2.1 will be presented as well.

Table 16 shows the corpus statistics of the RWTH-PHOENIX-v2.1 translation corpus. The corpus was updated in February 2011 and contains 1146 more sentences than the previous snapshot RWTH-PHOENIX-v2.0. Note that the statistics differ from the statistics for sign language recognition. The reason is that the videos in RWTH-PHOENIX-v1.0 had a different design, showing the interpreter as an inscreen-video and warping this image to portray a 3D effect. These videos are not used in the current recognition framework because of this different setup. However, since this change in the screen setup does not affect the glosses or the German text, we still decided to use this part of the annotated corpus to increase the amount of training data for WP5.

In statistical machine translation of spoken languages, a part of the training data is withheld to optimize weights of the feature functions so that the translation system generalizes well to unseen data. (In Figure 2, we refer to this standard method as the *baseline*.) On such large-scale tasks, a development set of 500 sentences usually makes up only a negligible portion of the training data. On the other hand, for our current corpora,

¹<http://code.google.com/p/morphisto/>

Table 16: Corpus statistics for the RWTH-PHOENIX-v2.1 translation corpus

		Glosses	German
Train:	Sentences	3711	
	Running Words	35 480	48 897
	Vocabulary	997	1 799
	Type/Token ratio	35.6	27.2
	Singletons / Voc	34.9%	36.9%
Test:	Sentences	512	
	Running Words	5 053	7 170
	OOVs (running)	1.3%	(1.6%)
	Trigram ppl.	40.7	18.0

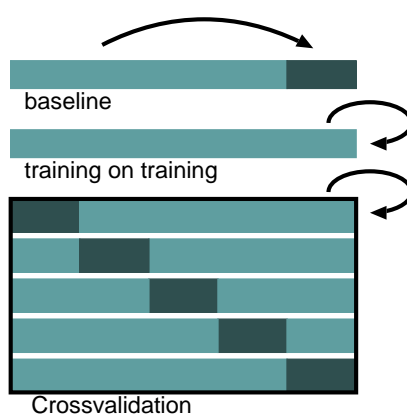


Figure 2: Graphical representation of the different optimization methods

holding back a development set of the same size strips away 20% of the training material. In [9], the authors claim that the best way to optimize the scaling factors on their corpus is to train them on the complete training set, thus not utilizing a development set at all. This approach, which we will denote as *training-on-training*, obviously bears the danger of over-fitting. We therefore utilize a method similar to *cross-validation*: We train five translation system, each time holding back a portion of one fifth of the training corpus. In each optimization iteration, we concatenate the n -best lists of all the individual systems for a complete training set translation. For a visual representation of the different optimization methods see Figure 2. The described method led to some improvements over the baseline. For more details see [15].

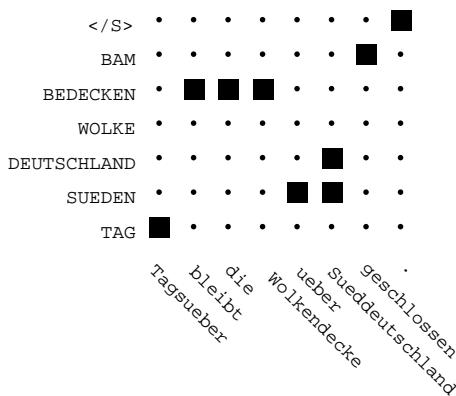
In spoken German, noun compounds are written as a single word. Since these compounds usually occur less frequent than the individual nouns they are made up of, it is advisable to apply techniques to split them, because a translation system might know the individual nouns, but not the noun compound. In our case, we split noun compounds to improve the automatic word alignment procedure, and we apply the morphological analyzer Morphisto to perform the splitting. Then we train the word alignment on the split corpus. However, since the translation output of such a system would contain split noun compounds, we did not choose to train the translation system on this corpus but instead merged the obtained alignment back such that it fits the original corpus. We refer to this technique as *crunched alignments*. See Figure 3 for an example of the method. The results in Table 17 show that this procedure leads to improvements both in BLEU and TER.

We also applied additional techniques well-established in the machine translation of spoken languages, such as syntactical techniques (syntactic labels + parsematch) for the hierarchical system and discriminative lexicon models (DWL + Triplet models) for the phrase-based system. For a detailed description of these techniques also see our publication [15].

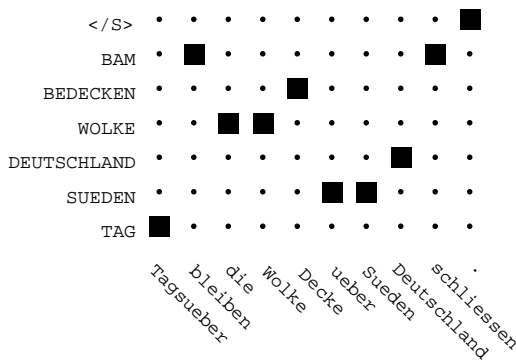
After training both a phrase-based and a hierarchical phrase-based system, we also performed a ROVER-like system combination [10] of these systems. The system combination again led to significant improvements over the individual systems.

Note that there is a significant improvement by switching from the smaller RWTH-PHOENIX-v2.0 to the newest snapshot RWTH-PHOENIX-v2.1. This indicates that our statistical methods strongly depend on sufficient training data, and that further improvements can be expected if more annotated data were available.

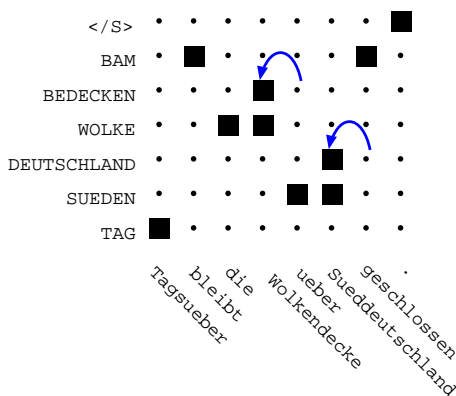
Since corpus-based statistical machine translation systems can be easily applied to new tasks by training



(a) Baseline alignment.



(b) Split alignment.



(c) Crunched alignment.

Figure 3: Example of the alignment crunching effect, taken from the RWTH-PHOENIX-v2.0 corpus, on the sentence “Tagsüber bleibt die Wolkendecke in Süddeutschland geschlossen” (Engl.: “During the course of the day, the cloud cover in southern Germany remains dense.”). The word “Wolkendecke” (“cloud cover”) is a singleton, but “Wolke” (“cloud”) is of course well-known and “Decke” is known from “Schneedecke” (“snow cover”). Thus, in (a) the alignment has errors, but for compound split German in (b) the quality is much better. After crunching the alignment in (c), the alignment structure matches the original German sentence.

Table 17: Translation results on RWTH-PHOENIX Glosses \Rightarrow Spoken German

Decoder	System	v2.0		v2.1	
		BLEU	TER	BLEU	TER
Hierarchical System	Union Alignment	24.3	67.3	25.3	66.3
	Crunched Alignment	25.0	66.5	26.2	65.3
	Syntactic labels + Parsematch	24.0	68.1	25.1	67.0
Phrase-based System	Intersection Alignment	24.6	65.1	25.8	63.8
	Crunched Alignment	25.1	64.2	26.5	64.1
	DWL + Triplet	24.6	64.1	27.3	63.2
Both	System Combination	26.0	63.8	28.1	62.7

Table 18: Translation results on RWTH-PHOENIX-v2.0: Spoken German \Rightarrow Glosses
Phrase-based system

	BLEU	TER
baseline	14.2	79.3
+ categories	15.4	77.0
+ compound splitting	15.9	75.8
+ advanced lexicon models	16.3	74.3
Hierarchical system		
	BLEU	TER
baseline (incl. categories)	15.5	84.5
+ sentence end markers	16.3	76.1
+ compound splitting	16.1	74.1

the system on a different corpus, we also applied our systems for the translation of the opposite translation direction, i.e. from spoken German to the glosses. In our automatized framework, this hardly required any effort and only some computational power. The results are shown in Table 18. The performance in this translation direction is clearly worse than in the direction from Glosses to the spoken language. One indication that it is more difficult to translate into glosses is the higher trigram perplexity on the gloss side (see Table 16).

2.3.2.2 Task 5.2 Handling translation of parallel information channels

In the course of reimplementing its phrase-based decoder in the open source framework Jane, RWTH is currently implementing word graph input to the decoder. With such input, it is possible to process parallel information channels. Besides processing information that is annotated on different tiers (c.f. Task 5.5), it is also possible to use word-graph output of a sign language recognition system to take into account multiple suggestions made by the recognition system instead of only considering the best hypothesis. When translating such a gloss-graph into the spoken language, the translation system can thus also choose glosses which were not chosen in the best path of the recognition system. We expect the implementation to be finished in the next months.

2.3.2.3 Task 5.3 Handling of spatial reference points

In sign languages, entities can be 'stored' in the signing space for later reference. In the case of the RWTH-PHOENIX-v2.0 corpus, references are often used to point out geographical locations. When referring back to a location, pointing gestures or signs which incorporate positions are used. In the gloss annotation, pointing gestures are enriched by information about the location which was referred to. Since the translations in Task 5.1 were performed on the annotations, they include this location information. To estimate the impact of this information on translation quality, we performed a first preliminary experiment by leaving out this additional information concerning locations. The results can be seen in Table 19. The comparison shows that the system significantly deteriorates if not provided with this spatial information. However, the deterioration is less drastic than we expected. We are currently working on the incorporation of tracking information provided by WP4 into

the translation process to obtain such location information. For this, classification methods will be used to map the position information to distinct words.

Table 19: Estimating the impact of spatial information (RWTH-PHOENIX-v2.0 corpus)

	Phrase-Based	
	BLEU	TER
without spatial information	22.8	65.6
with spatial information	24.1	64.8

2.3.2.4 Task 5.4: Handling of temporal signs and incorporations

Currently, signs modified by incorporation are treated as a whole by the recognition system, e.g. in the case of German Sign Language, the incorporation of the horizontal movement indicating “week” into the sign “three”, is recognized as one sign “THREE-WEEKS”. With the current annotation information given in the corpora, a distinction between such subelements of a sign is not possible. For this, a lexicon which maps signs to such subunits needs to be created manually.

The distinction of such subelements is best represented as a parallel input stream, since a representation on one input stream would make this approach almost identical to the recognition of incorporated signs recognized as a whole. Since our implementation of parallel input streams is not finished yet, we have not proceeded further in this task yet, but are planning to do so as soon as the processing of such streams is possible.

2.3.2.5 Task 5.5: Comparing parallel input channel approaches for multi-modal translation

In the Corpus-NGT, glosses are annotated independently for each hand, and headshake annotation is provided on an additional tier as well. We therefore have to address the problem of parallel input channels. While certainly more accurate, this annotation procedure presents a challenge for translation systems, which usually only permit one input stream. In the following sections, we first address the issue of the two hands being glossed individually. Then, we discuss first experiments on including headshake information into our system.

The example in Figure 4(a) shows that for some sentences, the dominant hand covers all words of the sentence and the non-dominant hand remains motionless for signs that only require one active hand. However, this is not always the case. The example in Figure 4(b) shows the transcription of a signer who switches the active signing hand within one sentence.

We performed three experiments. First, we only employed the *right hand* information as our source input data and define this as our baseline. A next approximation is to select for each sentence the glosses of the hand that signs more words, an approach which we call *active hand*. In a third step, we parsed the annotation file again, matched the timing of the individual glosses, and time-aligned both gloss tiers, omitting word duplications whenever both hands sign the same (*merged hands*). Note that this method still does not capture the whole expressiveness of sign languages, e.g. a signer might sign “NEWSPAPER” with both hands, keeping the non-dominant hand in this position but signing “COFFEE” with his dominant hand in the meantime, a signed construction which could be translated as “drinking coffee while reading the newspaper”.

The results can be found in Table 22. Switching from the right hand to the active hand gives a significant improvement of 8.1 in the TER score, and the merged hand approach further improves the BLEU score by 2.2.

When running the experiments to include headshake annotation into our MT system, we switched to the latest snapshot of the Corpus-NGT. Note however that in the process of updating the corpus, the annotation and the translations throughout the whole corpus were altered, which includes the test data. This makes the two experiments non-comparable. Corpus-NGT contains several different symbols for headshakes. A summary can be found in Table 20. The annotation distinguishes between headshakes indicating a negation and those not indicating a negation. As in the previous section, we included headshakes by merging them with the two tiers of the manual glosses by sorting them according to the timeline. This led to several problems, for example in the case when a person shakes his head continuously during several consecutive sentences. In this case, headshakes were added to each sentence.

We compared four setups: the first contains no headshake annotation, forming the baseline. Here, the difference between the newest snapshot and the older snapshot of Corpus-NGT is visible. While the BLEU degraded, the TER significantly improved in the newer version. The second system contains all the different kinds of headshake symbols. In the third system, only the most common symbol “N” was included, whereas in the last system, all headshakes indicating negation were mapped onto one symbol, while headshakes not

indicating negation were not included. The results in Table 21 show some improvement of the last setup with respect to BLEU, but a degradation in TER. Probably, more sophisticated methods to include headshakes are necessary instead of the simple merging strategy we applied.

Table 20: Different annotation symbols for headshakes in Corpus-NGT

Code	Negation	Description	Frequency
N	✓	cooc. with manual signs	690
Nx	✓	falling in between two manual signs	116
Nf	✓	not cooc.with manual signs (i.e. especially of the addressee)	287
Nn	✗	cooc. with manual signs	22
Ns	✓	Head sway from side to side	35
Nsx	✗	Head sway from side to side	22

Table 21: Results of translation systems including different variants of headshake annotation

	BLEU	TER
No headshake	8.9	70.9
All headshakes	8.1	78.7
Only "N"	9.2	71.8
Map negations to one symbol	9.7	71.6

While in general the error measures on Corpus-NGT are still unsatisfactory, we expect to gain better overall results with more training data, which will soon be released, and consider these results as a first waystage for SLMT in a broader domain. Moreover, we expect better results, since the new data will also include multiple reference translations, which stabilizes the optimization procedure and measures the translation quality more accurately.

Currently, we are working on an implementation to deal with multiple input streams in parallel. This approach is however more complex than the so-called factored translation model known in spoken language translations [8], since information provided by different modalities in parallel (manual gestures, facial expression, movement of head and gaze, etc.) might be translated in different parts of the spoken sentence, that is, the system needs to translate glosses on different tiers at different times.

right hand	MOEILJK DOEN OVER	COMMUNICEREN PO	MET	IX HOREND	MENSEN PO
left hand	MOEILJK DOEN	COMMUNICEREN	MET		MENSEN
Dutch	Erg veel moeite doet om te communiceren met horende mensen.				
	(a) "It is quite hard to communicate with hearing persons."				
right hand	ALS IX-1 LANG NIET		BETEKENEN EMAIL	BETEKENEN	GEBAREN IX-1 PO
left hand		NIET HEEN CLUBHUIS TOE	BETEKENEN		GEBAREN IX-1 PO
Dutch	als je lang niet naar het clubhuis gaat , weet je het gebaar voor het woord e-mail bijvoorbeeld niet .				
	(b) "If you haven't been to the club house for some time, you will not know the sign for the word 'email'. "				

Figure 4: Example sentences from the Corpus-NGT corpus. Each hand is annotated on a separated tier.

2.3.3 Clearly significant results

RWTH has significantly improved its translation system over the last year by developing and applying several methods which are specifically suitable for sign language corpora. First of all, a method similar to crossvalidation was applied to stabilize the parameter optimization on sign language corpora which are rather small when compared to corpora in spoken language translation. Moreover, by applying a morphological analyzer, linguistic

Table 22: Corpus-NGT Translation results applying different strategies for using two gloss streams

	Phrase-Based		Hierarchical	
	BLEU	TER	BLEU	TER
Right Hand	6.8	86.7	8.1	79.2
Active Hand	8.5	78.6	8.3	77.8
Merged Hands	10.7	78.3	10.2	77.4

knowledge about the spoken languages were taken into account to improve the alignment quality. On Corpus-NGT, a first and simple method was developed to deal with parallel input streams by merging these streams according to the timeline. These methods helped to improve the translation quality mainly in the translation direction from sign language into the spoken language, but they can also be applied in systems in the opposite translation direction.

3 Objectives for the next Evaluation

The next evaluation will be delivered on month 35 (end of February 2012). The evaluation will be carried out over the final prototypes generated during the third year of the project: final prototypes for multimodal visual analysis (D.3.5), extended prototypes for sign language recognition (D.4.3) and extended prototypes for sign language translation (D5.3).

4 References

- [1] Onno Crasborn, Inge Zwitserlood, and Johan Ros. Corpus-ngt. an open access digital corpus of movies with annotations of sign language of the netherlands. Technical report, Centre for Language Studies, Radboud University Nijmegen, 2008. <http://www.corpusngt.nl>.
- [2] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking using dynamic programming for appearance-based sign language recognition. In *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, pages 293–298, Southampton, April 2006.
- [3] Philippe Dreuw, Jens Forster, Thomas Deselaers, and Hermann Ney. Efficient approximations to model-based joint tracking and recognition of continuous sign language. In *IEEE International Conference Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, September 2008.
- [4] Philippe Dreuw, Jens Forster, Yannick Gweth, Daniel Stein, Hermann Ney, Gregorio Martinez, Jaume Verges Llahi, Onno Crasborn, Ellen Ormel, Wei Du, Thomas Hoyoux, Justus Piater, Jose Miguel Moya Lazaro, and Mark Wheatley. Signspeak – understanding, recognition, and translation of sign languages. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Malta, May 2010.
- [5] Philippe Dreuw, Hermann Ney, Gregorio Martinez, Onno Crasborn, Justus Piater, Jose Miguel Moya, and Mark Wheatley. The signspeak project - bridging the gap between signers and speakers. In *International Conference on Language Resources and Evaluation*, Valletta, Malta, May 2010.
- [6] Jens Forster, Daniel Stein, Ellen Ormel, Onno Crasborn, and Hermann Ney. Best practice for sign language data collections regarding the needs of data-driven recognition and translation. In *4th LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, Malta, May 2010.
- [7] Stephan Kanthak, Achim Sixtus, Sirko Molau, Ralf Schlüter, and Hermann Ney. *Fast Search for Large Vocabulary Speech Recognition*, chapter "From Speech Input to Augmented Word Lattices", pages 63–78. Springer Verlag, Berlin, Heidelberg, New York, July 2000.
- [8] P. Koehn and H. Hoang. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, volume 868, page 876, 2007.

- [9] Guillem Massó and Toni Badia. Dealing with Sign Language Morphemes for Statistical Machine Translation. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 154–157, Valletta, Malta, May 2010.
- [10] Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September 2008.
- [11] Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, and Hermann Ney. The RWTH Statistical Machine Translation System for the IWSLT 2006 evaluation. In *IWSLT*, pages 103–110, Kyoto, Japan, November 2006. Best Paper Award.
- [12] Justus Piater, Thomas Hoyoux, and Wei Du. Video analysis for continuous sign language recognition. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 192–195, Valletta, Malta, May 2010.
- [13] D. Rybach. Appearance-based features for automatic continuous sign language recognition. Diploma thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany, June 2006.
- [14] D. Stein, J. Bungeroth, and H. Ney. Morpho-Syntax Based Statistical Methods for Sign Language Translation. In *11th EAMT*, pages 169–177, Oslo, Norway, June 2006.
- [15] Daniel Stein, Christoph Schmidt, and Hermann Ney. Sign Language Machine Translation Overkill. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *International Workshop on Spoken Language Translation*, pages 337–344, Paris, France, December 2010.
- [16] David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July 2010.
- [17] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.
- [18] Andras Zolnay. *Acoustic Feature Combination for Speech Recognition*. PhD thesis, RWTH Aachen University, Aachen, Germany, August 2006.

Acknowledgments.

This work received funding from the European Community's Seventh Framework Programme under grant agreement number 231424 (FP7-ICT-2007-3).