Joint Lemmatization and Morphological Tagging with LEMMING
Appendix

Section 1 and 2 illustrate the edit tree extraction and application by Chrupała (2008) and also provide pseudo code. Section 3 and 4 provide an extended overview of our results.

1 Tree Extraction

function TREE(x,y)
  $i_s, i_e, j_s, j_e \leftarrow LCS(x, y)$
  if $i_e - i_s = 0$ **then**
    return SUB(x,y)
  else
    return (TREE($x_0^{i_s}, y_0^{j_s}$), i_s, TREE($x_{i_e}^{|x|}, y_{j_e}^{|y|}$), $|x| - i_e$)

Create a tree given a form-lemma pair $\langle x, y \rangle$. LCS returns the start and end indexes of the LCS in x and y. $x_{i_s}^{i_e}$ denotes the substring of x starting at index i_s (inclusive) and ending at index i_e (exclusive). $i_e - i_s$ thus equals the length of this substring. $|x|$ denotes the length of x. Note that the tree does not store the LCS, but only the length of the prefix and suffix. This way the tree for *umgeschaut* can also be applied to transform *umgebaut* "renovated" into *umbauen* "to renovate".

For the example *umgeschaut-umschauen*, the LCS is the stem *schau*. The function then recursively transforms *umge* into *um* and *t* into *en*. The prefix and suffix lengths of the form are 4 and 1 respectively. The left sub-node needs to transform *umge* into *um*. The new LCS is *um*. The new prefix and suffix lengths are 0 and 2 respectively. As the new prefix is empty the is nothing more to do. The suffix node needs to transform *ge* into the empty string ϵ. As the new LCS of the suffix is empty, because *ge* and ϵ have no character in common, the node is represented as a substitution node. The remaining transformation *t* into *en* is also represented as a substitution, resulting in the tree in Figure 1:

2 Tree Application

function APPLY(tree, x)

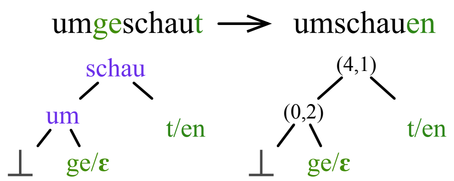

Figure 1: Edit tree for the inflected form *umgeschaut* "looked around" and its lemma *umschauen* "to look around". The right tree is the actual edit tree we use in our model, the left tree visualizes what each node corresponds to. Note how the root node stores the length of the prefix *umge* and the suffix *t*.

if tree is a LCS node **then**
  tree $\rightarrow$ $\text{tree}_i, i_l, \text{tree}_j, j_l$
  if $|x| < i_l + j_l$ **then** ▷ Prefix and Suffix do not fit.
    return $\perp$
  $p = \text{APPLY}(\text{tree}_i, x_0^{i_l})$ ▷ Create prefix.
  if p is $\perp$ **then** ▷ Prefix tree cannot be applied.
    return $\perp$
  $s = \text{APPLY}(\text{tree}_j, x_{|x|-j_l}^{|x|})$ ▷ Create suffix.
  if s is $\perp$ **then** ▷ Suffix tree cannot be applied.
    return $\perp$
  return p + $x_{i_l}^{|x|-j_l}$ + s ▷ Concatenate prefix, LCS and suffix.
else ▷ tree is a SUB node
  tree $\rightarrow u, v$
  if $x = u$ **then** ▷ If x and u match return v
    return v
  return $\perp$ ▷ tree cannot be applied.

In the code + represents string concatenation and $\perp$ a null string, meaning that the tree cannot be applied to the form. We first run the tree depicted in Figure 1 on the form *angebaut* “attached (to a building)”. The first node is a LCS node specifying that prefix and suffix should have length 4 and 1, respectively. We thus recursively apply the left child node to the prefix string *ange*. This is done by matching the length two prefix *an* and deleting *ge* yielding the intermediate result *anbaut*. We continue on the right side of the tree and replace *t* with *en*. This yield the final (correct) result *anbauen*. The application of the tree to the form *einbauen* “installed” would fail, as we would try to substitute a *ge* in *eing*.

3 Development Results

			cs		de		en		es		hu		la	
Baselines	SIMPLE MORFETTE	tag	86.08	69.83	79.05	56.02	94.72	87.95	96.29	87.21	94.37	85.20	83.86	57.77
		lemma	87.22	32.95	93.27	62.15	97.60	75.64	92.92	59.68	86.09	42.02	85.19	14.06
		joint	77.18	22.56	75.60	36.43	93.15	67.58	90.28	54.00	82.99	37.03	75.00	5.39
	SIMPLE PCRF	tag	89.82	76.83	85.05	65.22	**95.71**	**90.29**	96.83	89.00	**95.46**	88.17	86.35	65.21
		lemma	87.36	32.95	93.28	62.15	97.66	75.64	92.99	59.68	86.11	42.02	85.35	14.06
		joint	79.99	23.98	80.71	40.26	94.09	69.48	90.75	54.66	83.47	37.16	76.58	6.61
	JCK MORFETTE	tag	86.08	69.83	79.05	56.02	94.72	87.95	96.29	87.21	94.37	85.20	83.86	57.77
		lemma	96.24	82.59	97.67	88.80	98.71	92.50	97.61	86.76	97.48	91.16	93.26	63.09
		joint	84.32	61.67	78.10	51.42	94.43	84.61	94.67	78.39	93.15	80.73	81.34	43.71
	JCK PCRF	tag	89.82	76.83	85.05	65.22	**95.71**	**90.29**	96.83	89.00	**95.46**	88.17	86.35	65.21
		lemma	96.26	82.88	97.74	89.11	98.81	93.06	97.78	87.10	97.68	91.91	93.06	62.64
		joint	88.17	68.43	84.16	60.39	95.41	86.95	95.37	80.25	94.33	83.70	83.57	48.84
	MORFETTE MORFETTE	tag	86.08	69.83	79.05	56.02	94.72	87.95	96.29	87.21	94.37	85.20	83.86	57.77
		lemma	96.25	82.54	97.12	89.90	98.43	92.54	97.97	89.94	97.22	90.04	91.89	55.13
		joint	84.39	61.94	77.60	52.87	94.16	84.70	95.09	81.95	93.11	80.60	80.09	36.39
LEMMING	edit tree MORFETTE	tag	86.08	69.83	79.05	56.02	94.72	87.95	96.29	87.21	94.37	85.20	83.86	57.77
		lemma	96.29	82.93	97.84	89.78	98.71	92.63	97.91	89.00	97.31	90.49	93.00	61.68
		joint	84.45	62.36	78.32	52.75	94.43	84.79	94.97	80.68	93.08	80.48	80.92	41.27
	align MORFETTE	tag	86.08	69.83	79.05	56.02	94.72	87.95	96.29	87.21	94.37	85.20	83.86	57.77
		lemma	96.74	85.38	98.17	91.61	98.76	93.11	98.05	90.13	97.70	92.15	93.76	66.30
		joint	84.72	63.88	78.47	53.52	94.47	85.09	95.05	81.29	93.31	81.51	81.49	44.67
	dict MORFETTE	tag	86.08	69.83	79.05	56.02	94.72	87.95	96.29	87.21	94.37	85.20	83.86	57.77
		lemma	97.50	89.38	98.36	92.66	98.84	94.02	98.39	92.56	97.98	93.30	94.64	71.57
		joint	85.06	65.60	78.53	53.91	94.53	85.83	95.27	82.89	93.46	82.10	81.87	46.92
	morph MORFETTE	tag	86.08	69.83	79.05	56.02	NA	NA	96.29	87.21	94.37	85.20	83.86	57.77
		lemma	96.59	87.28	97.43	89.96	NA	NA	98.46	92.98	97.77	92.62	93.60	66.30
		joint	85.64	67.73	78.87	55.17	NA	NA	95.63	84.69	94.10	84.02	82.29	48.72
	edit tree PCRF	tag	89.82	76.83	85.05	65.22	**95.71**	**90.29**	96.83	89.00	**95.46**	88.17	86.35	65.21
		lemma	96.33	83.04	97.93	90.05	98.82	93.02	98.14	89.61	97.53	91.22	92.85	60.98
		joint	88.23	68.85	84.41	61.89	95.40	87.00	95.75	83.06	94.30	83.58	83.20	46.73
	align PCRF	tag	89.82	76.83	85.05	65.22	**95.71**	**90.29**	96.83	89.00	**95.46**	88.17	86.35	65.21
		lemma	96.80	85.72	98.24	91.94	98.88	93.71	98.27	90.75	97.91	92.91	93.50	64.89
		joint	88.57	70.69	84.59	62.95	95.45	87.47	95.81	83.64	94.52	84.60	83.83	50.32
	dict PCRF	tag	89.82	76.83	85.05	65.22	**95.71**	**90.29**	96.83	89.00	**95.46**	88.17	86.35	65.21
		lemma	97.62	89.88	98.43	92.95	98.94	**94.28**	98.62	93.23	98.23	94.16	94.49	70.60
		joint	88.97	72.79	84.63	63.16	**95.50**	**88.08**	96.02	85.24	94.73	85.46	84.41	53.79
	morph PCRF	tag	89.82	76.83	85.05	65.22	NA	NA	96.83	89.00	**95.46**	88.17	86.35	65.21
		lemma	97.38	89.55	98.24	92.42	NA	NA	**98.71**	93.97	98.30	94.38	94.39	70.54
		joint	89.30	74.32	84.81	64.08	NA	NA	96.17	86.29	95.15	86.87	84.62	55.07
	dict joint	tag	**90.47**	78.37	85.31	65.94	95.66	88.60	96.94	89.25	95.29	87.55	**86.77**	65.98
		lemma	98.34	92.96	98.66	94.07	**99.00**	94.23	98.68	94.08	98.53	95.46	**96.68**	**82.99**
		joint	89.82	75.47	84.94	64.19	95.48	86.52	96.11	85.43	94.71	85.40	85.85	60.98
	morph joint	tag	90.35	**79.61**	**85.52**	**67.69**	95.66	88.60	**96.95**	**89.55**	**95.46**	**88.85**	86.70	**66.82**
		lemma	**98.55**	**94.12**	**98.74**	**94.47**	**99.00**	94.23	98.68	**94.53**	**98.57**	**95.65**	96.51	82.22
		joint	**90.05**	**78.37**	**85.28**	**66.58**	95.48	86.52	**96.20**	**86.51**	**95.30**	**88.22**	**85.97**	**62.97**

Table A1: Development accuracies for the baselines and the different pipeline versions and a joint version of LEMMING. The numbers in each cell are general token accuracy and the token accuracy on unknown forms. Each cell specifies either a baseline ∈ {baseline, JCK, MORFETTE} or a LEMMING feature set ∈ {edit tree, align, dict, morph} and a tagger ∈ {MORFETTE, PCRF, joint}.

4 Test Results

			cs		de		en		es		hu		la	
Baselines	SIMPLE MORFETTE	tag	86.00	68.88	76.80	52.86	95.35	87.58	96.30	87.65	92.33	81.32	79.70	47.06
		lemma	86.81	31.16	91.42	58.68	97.88	78.95	92.86	59.34	84.67	38.95	83.60	20.43
		joint	76.96	21.22	72.83	33.27	93.82	68.98	90.14	53.45	80.41	33.08	71.62	7.52
	SIMPLE PCRF	tag	89.75	76.83	82.81	61.60	**96.45**	**90.68**	97.05	90.07	93.64	84.65	82.37	53.73
		lemma	86.94	31.16	91.48	58.68	97.92	78.95	92.92	59.34	84.73	38.95	83.80	20.43
		joint	79.66	22.86	77.78	36.49	94.87	71.62	90.81	54.98	81.06	33.27	73.51	8.58
	JCK MORFETTE	tag	86.00	68.88	76.80	52.86	95.35	87.58	96.30	87.65	92.33	81.32	79.70	47.06
		lemma	95.87	80.79	96.50	85.54	98.95	93.76	97.58	86.80	96.52	88.09	90.84	58.13
		joint	84.00	59.46	75.60	47.77	95.09	84.34	94.54	77.97	90.71	75.62	76.76	33.56
	JCK PCRF	tag	89.75	76.83	82.81	61.60	**96.45**	**90.68**	97.05	90.07	93.64	84.65	82.37	53.73
		lemma	95.95	81.28	96.63	85.84	99.08	94.28	97.69	87.19	96.69	88.66	90.79	58.23
		joint	87.85	67.00	81.60	55.97	96.17	87.32	95.44	80.62	92.15	78.89	79.51	39.07
	MORFETTE MORFETTE	tag	86.00	68.88	76.80	52.86	95.35	87.58	96.30	87.65	92.33	81.32	79.70	47.06
		lemma	95.88	80.70	95.88	87.55	98.73	93.56	98.02	90.16	96.06	86.21	90.08	54.16
		joint	84.10	59.92	74.87	49.22	94.89	84.44	95.03	81.68	90.60	75.11	76.57	32.50
LEMMING	edit tree MORFETTE	tag	86.00	68.88	76.80	52.86	95.35	87.58	96.30	87.65	92.33	81.32	79.70	47.06
		lemma	95.91	81.07	96.82	87.31	98.95	93.69	97.87	88.97	96.15	86.68	91.20	59.87
		joint	84.15	60.27	75.86	49.22	95.08	84.37	94.86	80.38	90.50	74.86	77.00	34.73
	align MORFETTE	tag	86.00	68.88	76.80	52.86	95.35	87.58	96.30	87.65	92.33	81.32	79.70	47.06
		lemma	96.43	83.83	97.22	89.25	98.99	94.15	98.04	90.28	96.87	89.72	91.64	62.36
		joint	84.47	61.97	76.05	50.07	95.11	84.73	94.98	81.30	90.94	76.76	77.18	35.79
	dict MORFETTE	tag	86.00	68.88	76.80	52.86	95.35	87.58	96.30	87.65	92.33	81.32	79.70	47.06
		lemma	97.24	88.13	97.58	91.03	99.07	95.06	98.32	92.36	97.28	91.32	92.97	69.14
		joint	84.88	64.14	76.13	50.48	95.19	85.55	95.19	82.91	91.14	77.48	77.52	37.59
	morph MORFETTE	tag	86.00	68.88	76.80	52.86	NA	NA	96.30	87.65	92.33	81.32	79.70	47.06
		lemma	96.52	86.69	96.57	88.26	NA	NA	98.53	93.73	96.70	89.30	91.53	62.31
		joint	85.57	66.73	76.56	51.90	NA	NA	95.63	85.02	91.92	79.73	78.18	39.65
	edit tree PCRF	tag	89.75	76.83	82.81	61.60	**96.45**	**90.68**	97.05	90.07	93.64	84.65	82.37	53.73
		lemma	96.05	81.68	96.99	87.69	99.08	94.25	98.00	89.56	96.38	87.54	91.37	60.88
		joint	88.03	68.04	81.96	57.91	96.17	87.38	95.80	83.42	91.93	78.15	79.75	40.34
	align PCRF	tag	89.75	76.83	82.81	61.60	**96.45**	**90.68**	97.05	90.07	93.64	84.65	82.37	53.73
		lemma	96.56	84.49	97.38	89.68	99.12	94.51	98.17	90.89	97.11	90.61	91.80	63.26
		joint	88.41	70.03	82.20	59.02	96.19	87.61	95.94	84.38	92.41	80.14	80.03	41.87
	dict PCRF	tag	89.75	76.83	82.81	61.60	**96.45**	**90.68**	97.05	90.07	93.64	84.65	82.37	53.73
		lemma	97.46	89.14	97.70	91.27	**99.21**	**95.59**	98.48	92.98	97.53	92.10	93.07	69.83
		joint	88.86	72.51	82.27	59.42	**96.27**	**88.49**	96.12	85.80	92.59	80.77	80.49	44.26
	morph PCRF	tag	89.75	76.83	82.81	61.60	NA	NA	97.05	90.07	93.64	84.65	82.37	53.73
		lemma	97.29	88.98	97.51	90.85	NA	NA	98.68	94.32	97.53	92.15	92.54	67.81
		joint	89.23	74.24	82.49	60.42	NA	NA	96.35	87.25	93.11	82.56	80.67	45.21
	dict joint	tag	**90.34**	78.47	**83.10**	62.36	96.32	89.70	97.11	90.13	93.64	84.78	82.89	54.69
		lemma	98.27	92.67	**98.10**	92.79	**99.21**	95.23	98.67	94.07	98.02	94.15	**95.58**	**81.74**
		joint	89.69	75.44	82.64	60.49	96.17	87.87	96.23	86.19	92.84	81.89	81.92	49.97
	morph joint	tag	90.20	**79.72**	**83.10**	**63.10**	96.32	89.70	**97.17**	**90.66**	**93.67**	**85.12**	**83.49**	**58.76**
		lemma	**98.42**	**93.46**	**98.10**	**93.02**	**99.21**	95.23	**98.78**	**94.86**	**98.08**	**94.26**	95.36	80.94
		joint	**89.90**	**78.34**	**82.84**	**62.10**	96.17	87.87	**96.41**	**87.47**	**93.40**	**84.15**	**82.57**	**54.63**

Table A2: Test accuracies for the baselines and the different pipeline versions and a joint version of LEMMING. The numbers in each cell are general token accuracy and the token accuracy on unknown forms. Each cell specifies either a baseline ∈ {baseline, JCK, MORFETTE} or a LEMMING feature set ∈ {edit tree, align, dict, morph} and a tagger ∈ {MORFETTE, PCRF, joint}.

	cs	de	es	hu	la
+dict	98.35	99.04	98.92	98.83	96.14
+morph	**99.03**	**99.47**	**99.09**	**99.41**	**96.73**

Table A3: Development accuracies for LEMMING with and without morphological attributes using *gold* tags.

	train				dev				test			
	sent	token	pos	morph	sent	token	form unk	lemma unk	sent	token	form unk	lemma unk
cs	5979	100012	12	266	5228	87988	19.86	9.79	4213	70348	19.89	9.73
de	5662	100009	51	204	5000	76704	17.11	13.53	5000	92004	19.51	15.64
en	4028	100012	46		1336	32092	8.06	6.16	1640	39590	8.52	6.56
es	3431	100027	12	226	1655	50368	13.37	8.85	1725	50630	13.41	9.16
hu	4390	100014	22	572	1051	29989	23.51	14.36	1009	19908	24.80	14.64
la	7122	59992	23	474	890	9475	17.59	6.43	891	9922	19.82	7.56

Table A4: Dataset statistics. Showing number of sentences (sent), tokens (token), POS tags (pos), morphological tags (morph) and token-based unknown form (form unk) and lemma (lemma unk) rates.

References

Grzegorz Chrupała. *Towards a machine-learning architecture for lexical functional grammar parsing*. PhD thesis, Dublin City University, 2008.