Reducing perplexities

(and word error rates)

with interpolation factors

CLS | Centre for Language Studies
Radboud University

Reducing perplexities and word error rates by improving the interpolation factors of a Bayesian skipgram language model

Louis Onrust
Centre for Language Studies, Radboud University
Center for Processing Speech and Images, KU Leuven

l.onrust@let.ru.nl
github.com/naiaden

HPYPLMs

Hierarchical Pitman-Yor process language model

- Bayesian language model
- HPYPLM follows the analogy of the Chinese restaurant process
- Generalisation of interpolated Kneser-Ney, which follows the analogy of a soup kitchen

Relevancy

- As interpolation term for sota language models
- First pass decoder ASR

We report ...

- effect of domains in training and testing
- perplexities
- word error rates

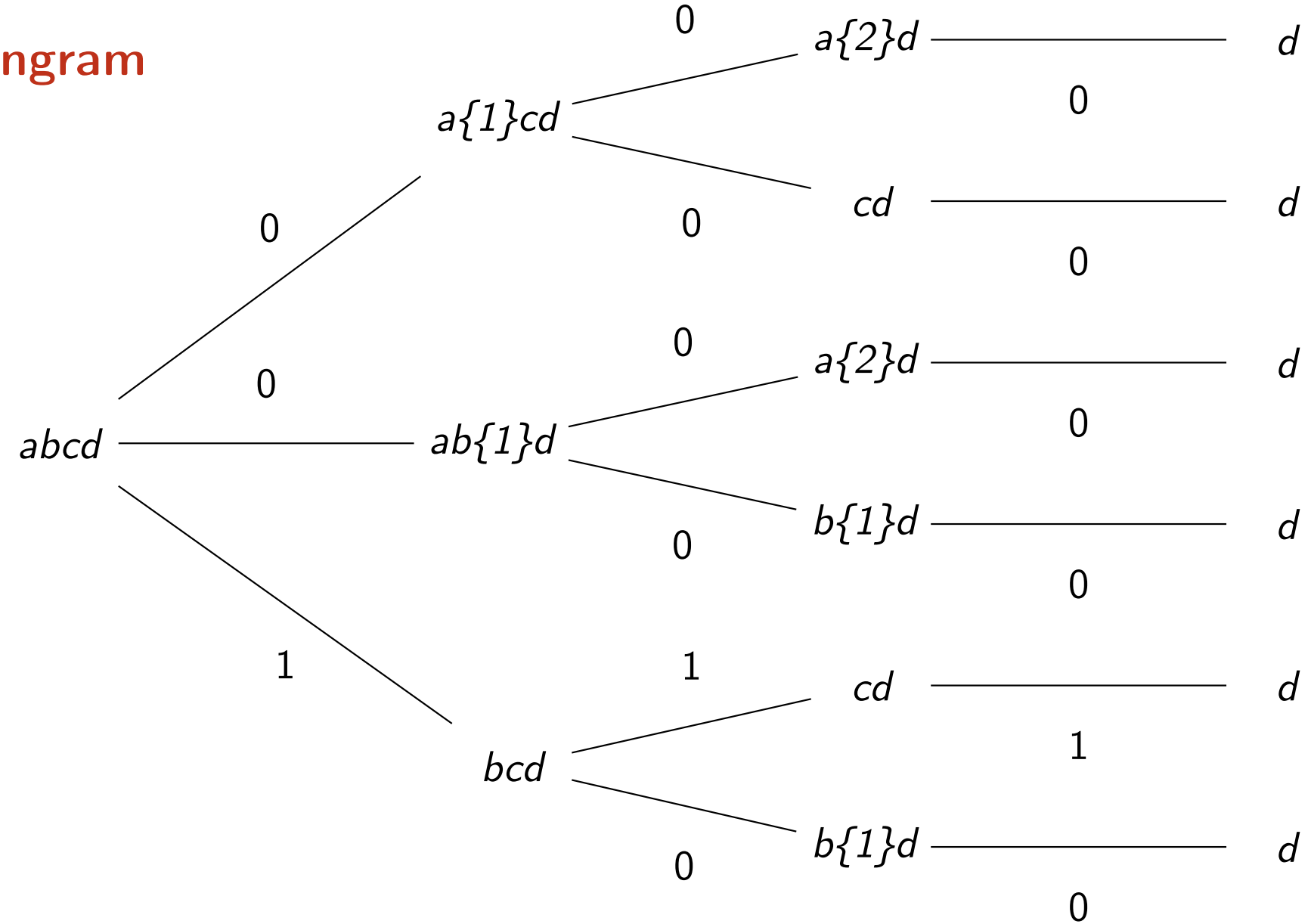
Backoff strategies

ngram
Only use the *n*-gram features

full
Additionally use skipgrams, full recursively

limited
Also skipgrams, but stop if pattern was in training material

ngram


abcd
0
0
1
a{1}cd
ab{1}d
bcd
0
0
0
0
1
0
a{2}d
cd
a{2}d
b{1}d
cd
b{1}d
0
0
0
0
1
0
d
d
d
d
d
d

Interpolation factors

uniform
All weights are uniformly distributed

npref
Give *n*-grams a higher weight

value
Each backoff step has a specific weight

count
Weights are determined by their context count

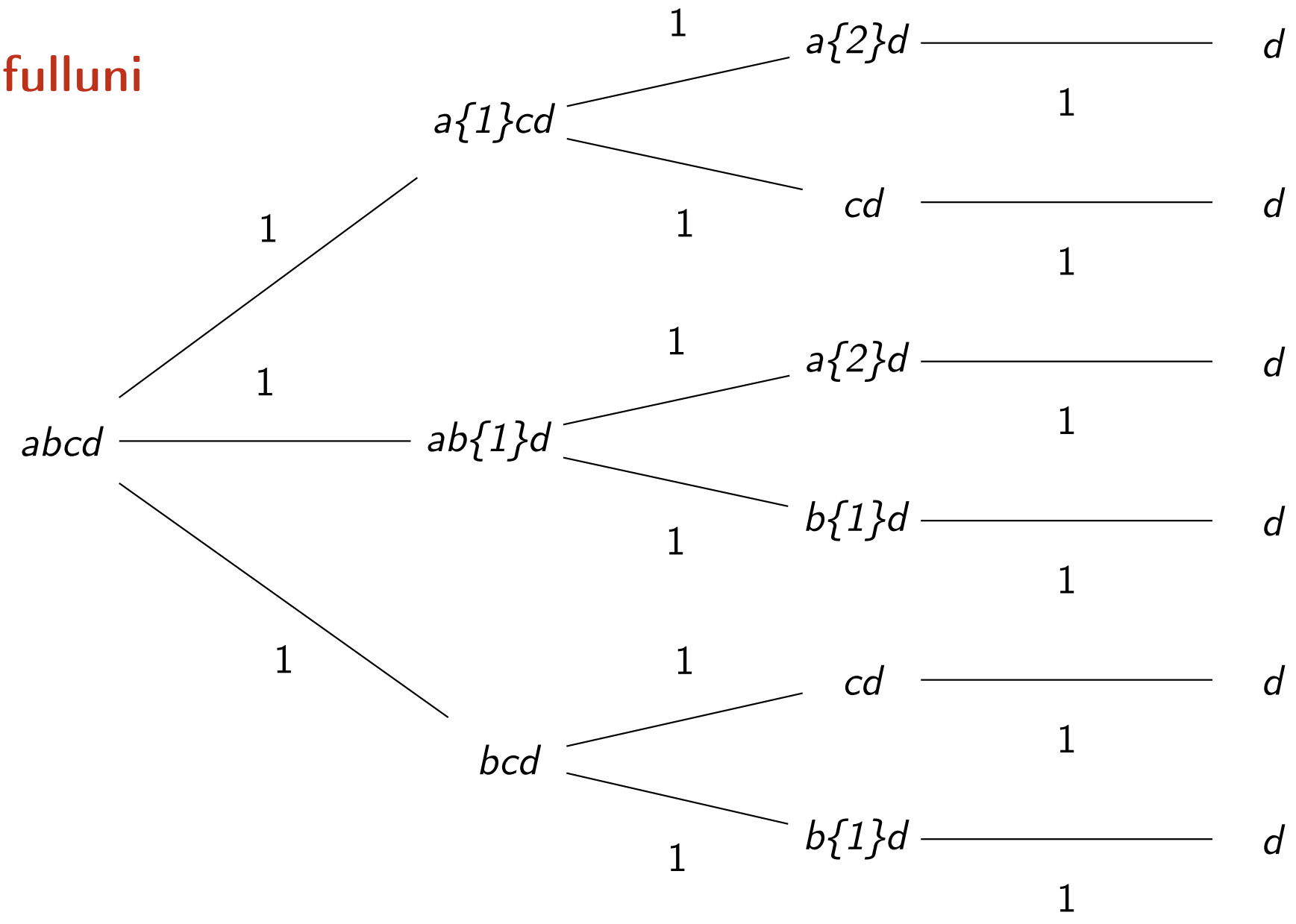
perplexity
Weights are determined by their context perplexity

entropy
Weights are determined by their context entropy

mle
Weights are determined by their maximum likelihood estimate

random
Weights are determined randomly

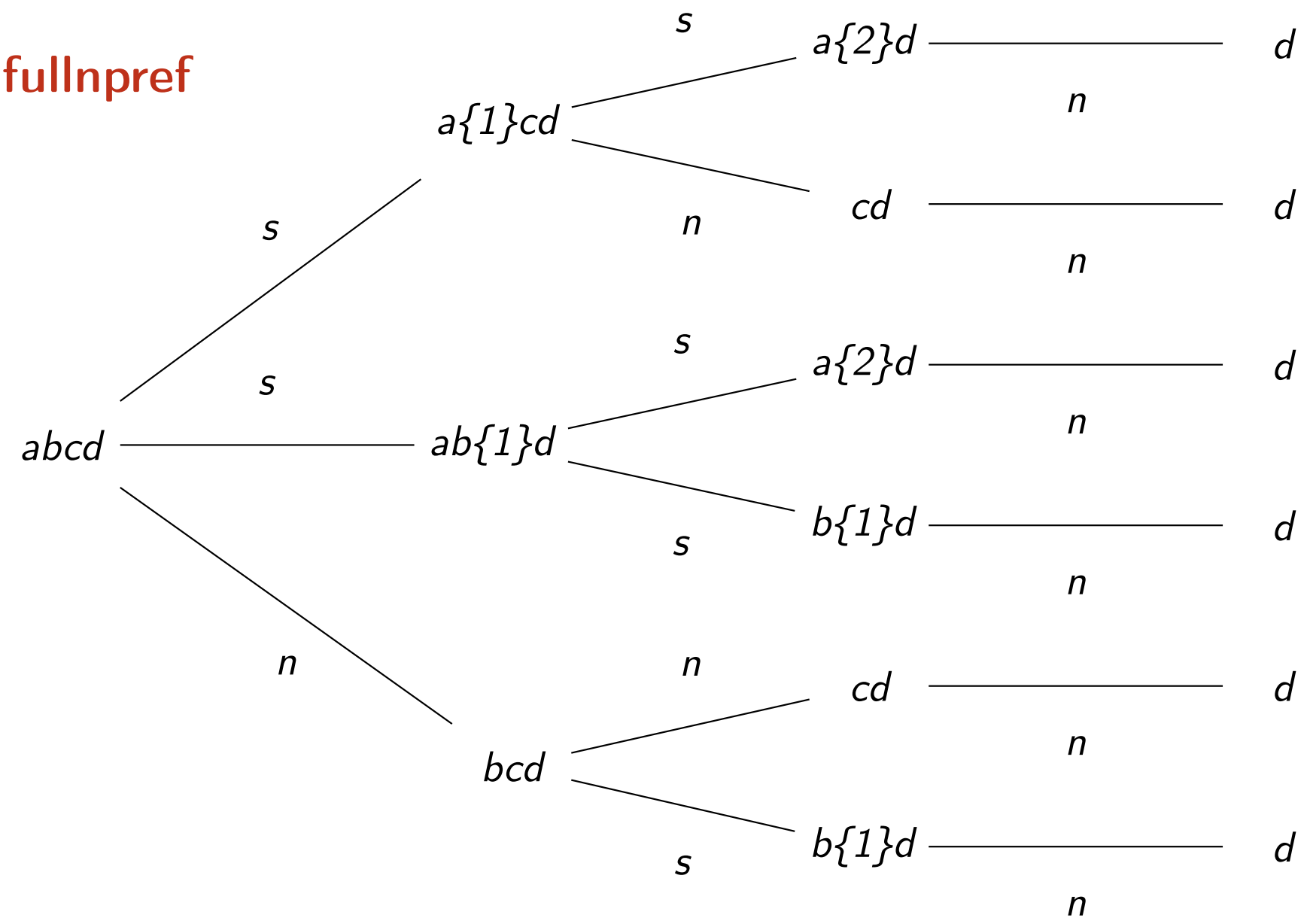
fulluni


abcd
1
1
1
a{1}cd
ab{1}d
bcd
1
1
1
1
1
1
a{2}d
cd
a{2}d
b{1}d
cd
b{1}d
1
1
1
1
1
1
d
d
d
d
d
d

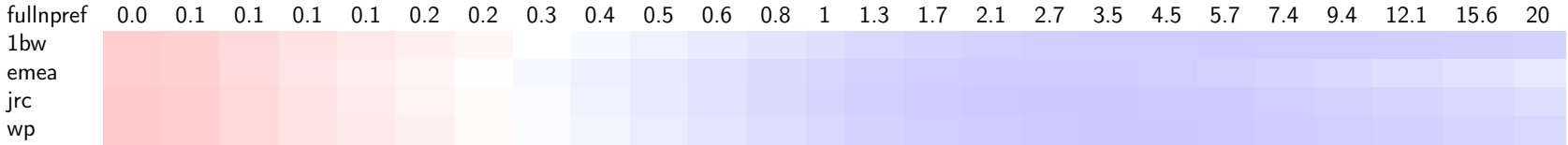
Perplexities on English data

training	1bw				emea					
test	1bw	emea	jrc	wp	1bw	emea	jrc	wp	1bw	
ngram	129	1 124	941	456	1 761	6	898	1 124	1 52	
uni	125	728	729	392	1 394	6	773	908	1 30	

fullnpref


abcd
s
s
n
a{1}cd
ab{1}d
bcd
s
n
s
n
s
n
a{2}d
cd
a{2}d
b{1}d
cd
b{1}d
n
n
n
n
n
n
d
d
d
d
d
d

Finding the *n*-gram skipgram weight ratio



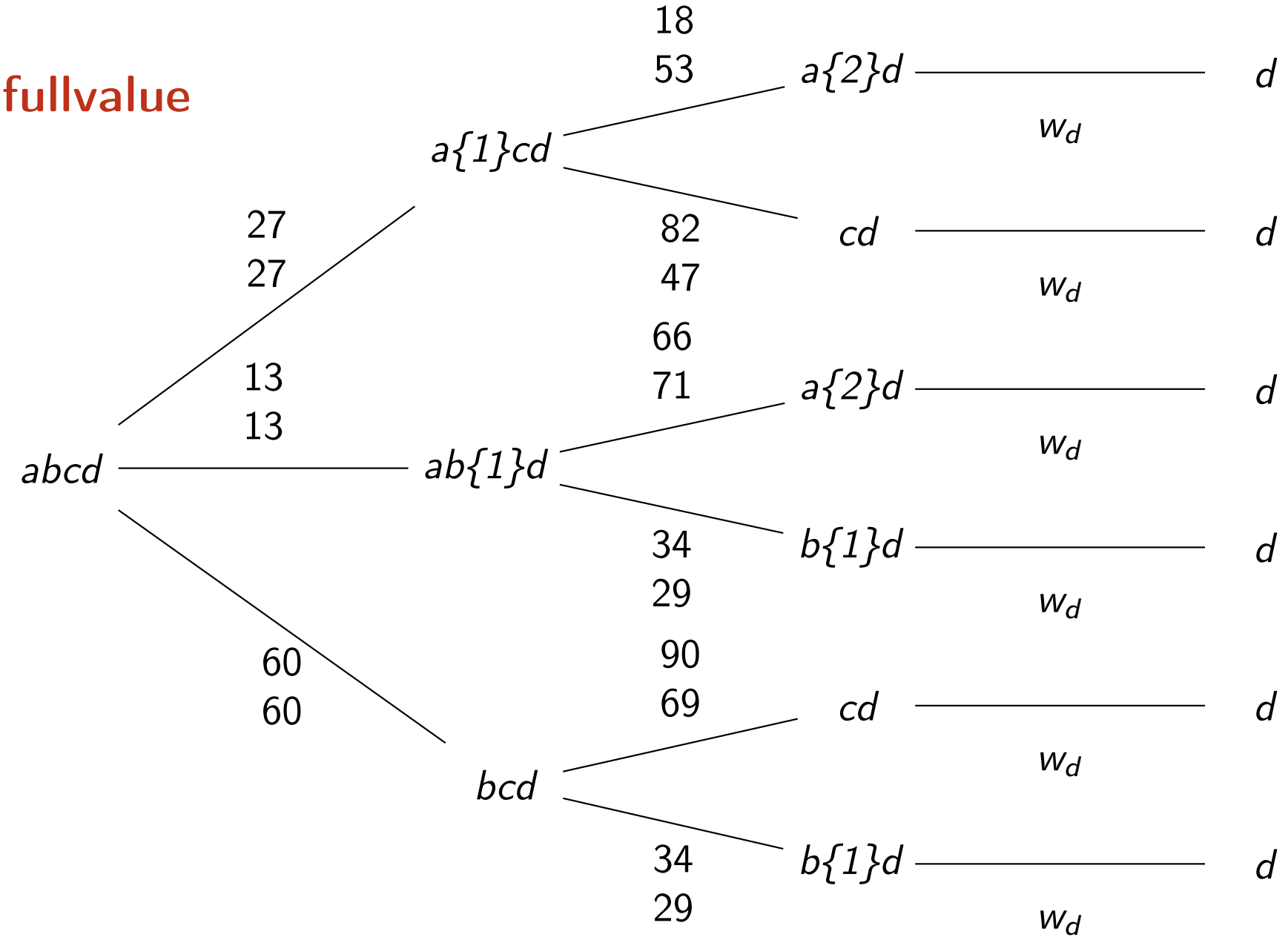
Perplexities on English data

training	1bw				emea				
test	1bw	emea	jrc	wp	1bw	emea	jrc	wp	1bw
ngram	129	1 124	941	456	1 761	6	898	1 124	1 52
uni	125	728	729	392	1 394	6	773	908	1 30
npref	118	700	694	372	1 306	6	705	853	1 21

Perplexities on English data

training	1bw				emea				
test	1bw	emea	jrc	wp	1bw	emea	jrc	wp	1
ngram	129	1 124	941	456	1 761	6	898	1 124	1
uni	125	728	729	392	1 394	6	773	908	1
npref	118	700	694	372	1 306	6	705	853	1
mle	125				1 931	6	1 015	1 225	1
count	122	893	885	421	1 681	6	889	1 075	1
ent	132	794	792	434	1 552	6	881	1 032	1
ppl	157	1 002	1 027	555	2 007	6	1 218	1 329	1

fullvalue


abcd
27
27
13
13
60
60
a{1}cd
ab{1}d
bcd
18
53
82
47
66
71
34
29
90
69
34
29
a{2}d
cd
a{2}d
b{1}d
cd
b{1}d
w_d
w_d
w_d
w_d
w_d
w_d
d
d
d
d
d
d

Perplexities on English data

training	1bw				emea			
test	1bw	emea	jrc	wp	1bw	emea	jrc	wp
ngram	129	1 124	941	456	1 761	6	898	1 124
uni	125	728	729	392	1 394	6	773	908
npref	118	700	694	372	1 306	6	705	853
mle	125				1 931	6	1 015	1 225
count	122	893	885	421	1 681	6	889	1 075
ent	132	794	792	434	1 552	6	881	1 032
ppl	157	1 002	1 027	555	2 007	6	1 218	1 329
1bw-value	115	713	694	366	1 212	6	655	655
emea-value	116	692	686	366	1 221	6	651	805
jrc-value	115	694	685	365	1 373	6	709	890
wp-value	115	696	685	317	1 212	6	654	654

Perplexities on English data

training	1bw				emea			
test	1bw	emea	jrc	wp	1bw	emea	jrc	wp
ngram	129	1 124	941	456	1 761	6	898	1 124
uni	125	728	729	392	1 394	6	773	908
npref	118	700	694	372	1 306	6	705	853
mle	125				1 931	6	1 015	1 225
count	122	893	885	421	1 681	6	889	1 075
ent	132	794	792	434	1 552	6	881	1 032
ppl	157	1 002	1 027	555	2 007	6	1 218	1 329
1bw-value	115	713	694	366	1 212	6	655	655
emea-value	116	692	686	366	1 221	6	651	805
jrc-value	115	694	685	365	1 373	6	709	890
wp-value	115	696	685	317	1 212	6	654	654
random	130	769	769	412	1 484	6	826	962

Perplexities on English data

training	1bw				emea			
test	1bw	emea	jrc	wp	1bw	emea	jrc	wp
ngram	129	1 124	941	456	1 761	6	898	1 124
limuni	134	759	756	406	1 422	6	793	926
limnpref	128	733	723	387	1 340	6	728	874
limcount	133	942	928	441	1 745	6	928	1 114
liment	144	832	825	453	1 583	6	904	1 053
limppl	172	1 055	1 075	851	2 049	6	1 252	1 359
limrandom	140	804	800	428	1 523	6	855	985

Perplexities on Dutch data

training	mediargus			
test	concat ref	concat ngram	utt ref	utt ngram
srilm	441	617	660	678
ngram	894	869	357	622
fulluni	724	818	305	600
limuni	777	882	326	647

WER on Dutch data

training	mediargus				
test	concat ref	concat ngram	utt ref	utt ngram	WER
srilm	441	617	660	678	33.03
ngram	894	869	357	622	32.92
fulluni	724	818	305	600	34.84
limuni	777	882	326	647	34.90