An Experimental Management System

Philipp Koehn
presented by Barry Haddow

16 September 2010

How do you run experiments?

Executing a Lot of Scripts

```
tokenize < corpus.en > corpus.en.tok
tokenize < corpus.fr > corpus.fr.tok
lowercase < corpus.en.tok > corpus.en.lc
lowercase < corpus.fr.tok > corpus.fr.lc
...
mert.perl ....
moses ...
mteval-v13.pl ...
```

Executing a Lot of Scripts

Oh wait, a mistake!

```
tokenize < corpus.en > corpus.en.tok
tokenize -l fr < corpus.fr > corpus.fr.tok
lowercase < corpus.en.tok > corpus.en.lc
lowercase < corpus.fr.tok > corpus.fr.lc
...
mert.perl ....
moses ...
mteval-v13.pl ...
```

Variations

- For instance, varying the distortion limit

```
for(my $dl=3;$dl<=10;$dl++) {
  ‘moses -dl $dl ... > output.dl-$dl‘;
  ‘wrap-xml.perl < output.dl-$dl > output.dl-$dl.sgm‘;
  ‘mteval-v13.pl -t output.dl-$dl ... > score.dl-$dl‘
}
```

- But:
  - needs to be customized for every case
  - what if some of the steps crash?
  - how schedule in parallel on different machines / cluster?

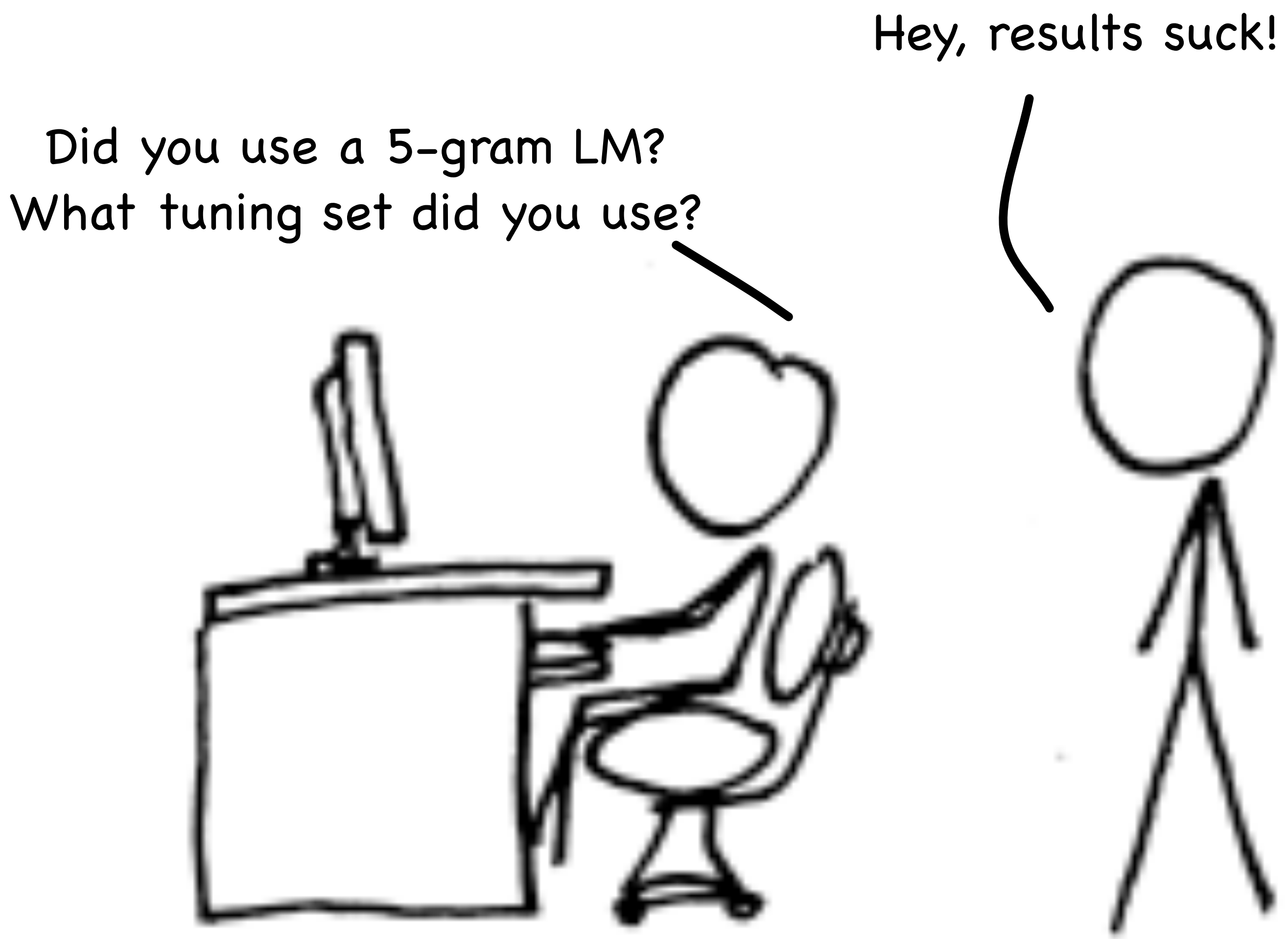
A New Student Arrives

A New Student Arrives

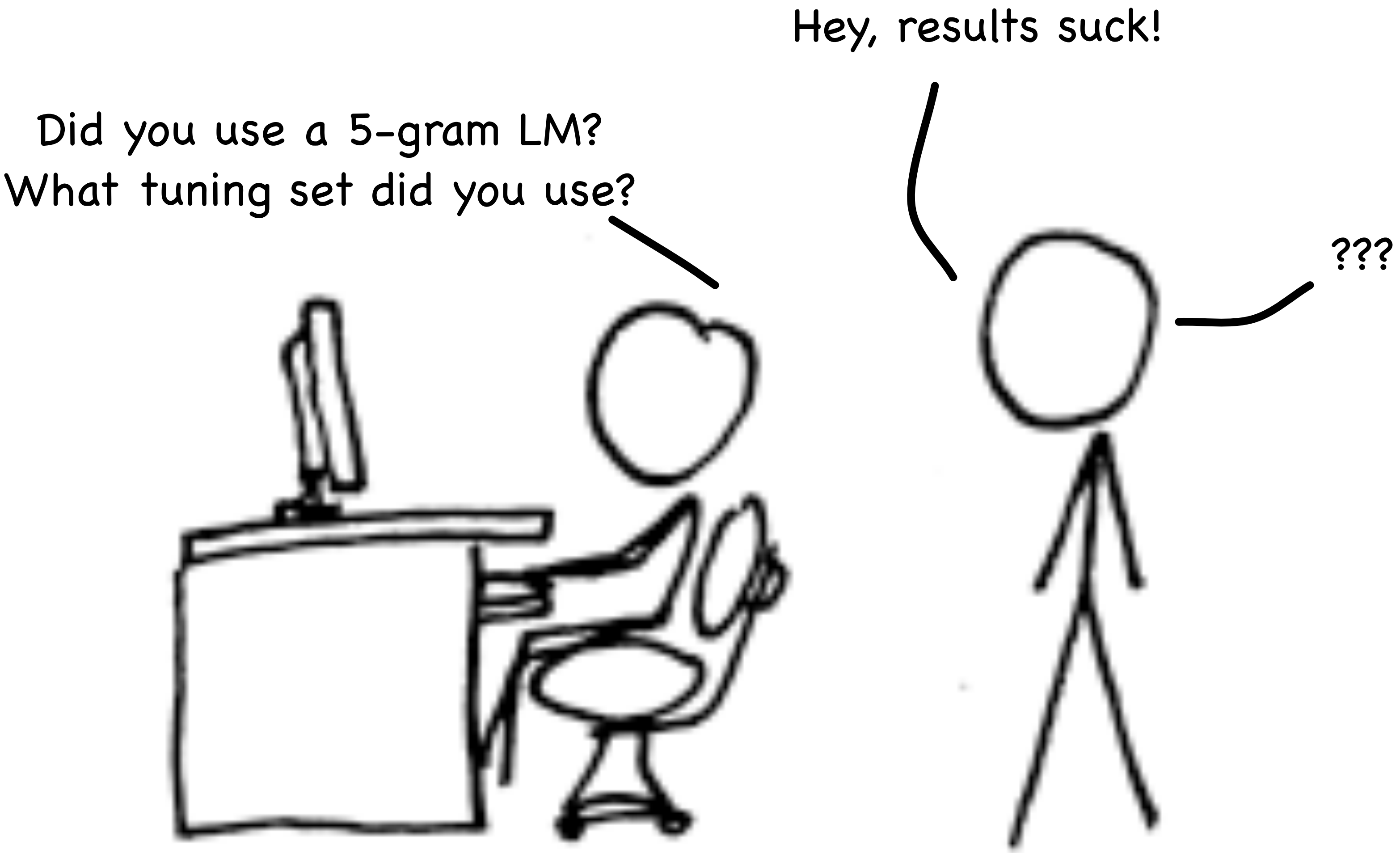
Hey, how do I build
a German-English system?
blah blah europarl blah
newstest2009 blah blah

One Week Later

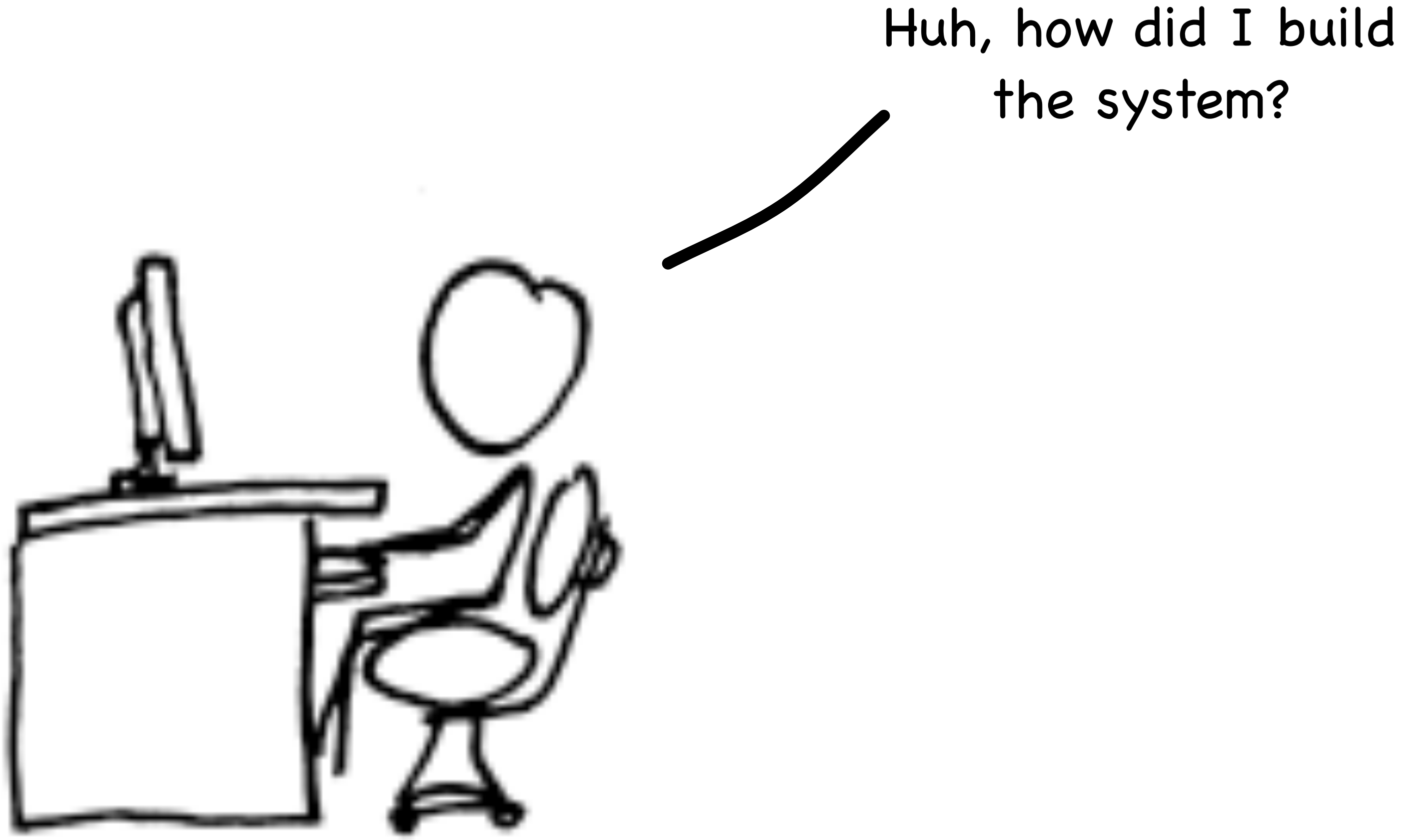
One Week Later



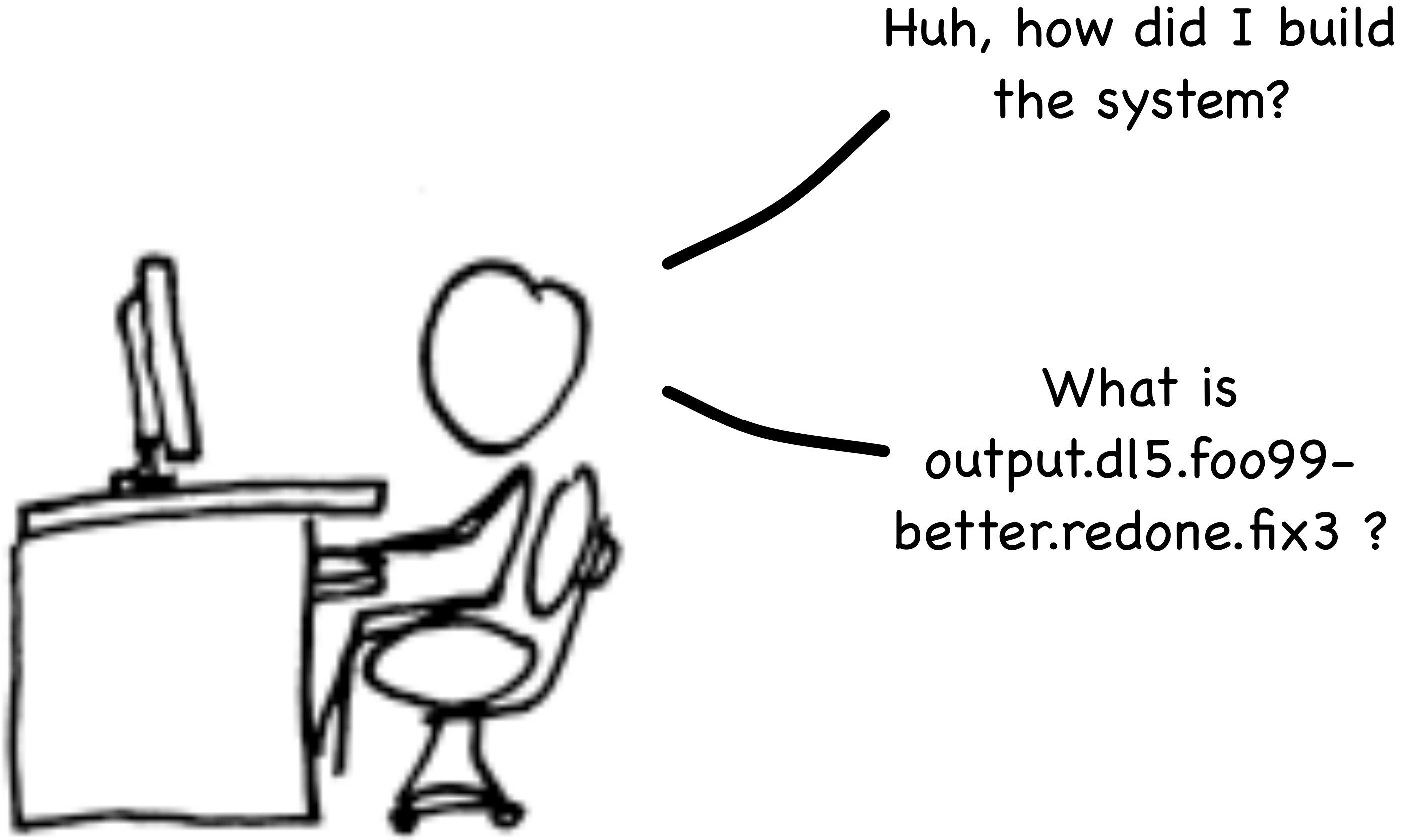
One Week Later



One Year Later



One Year Later


Huh, how did I build the system?
What is output.dl5.foo99-better.redone.fix3 ?

There must be a better way...

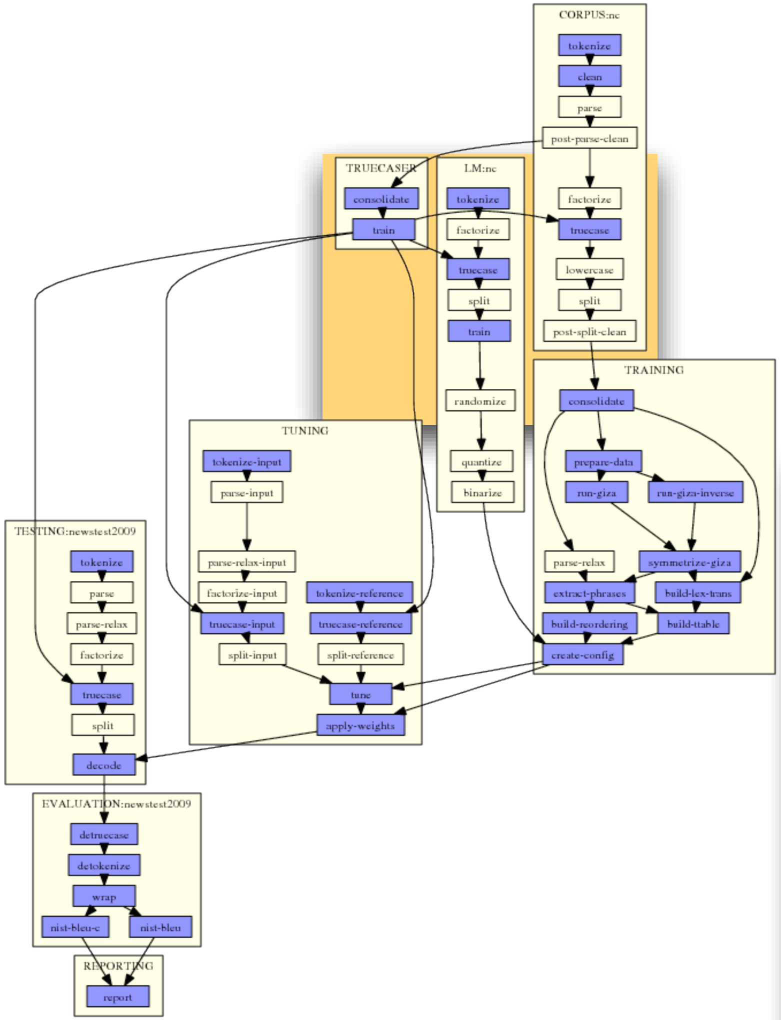
Experiment.perl

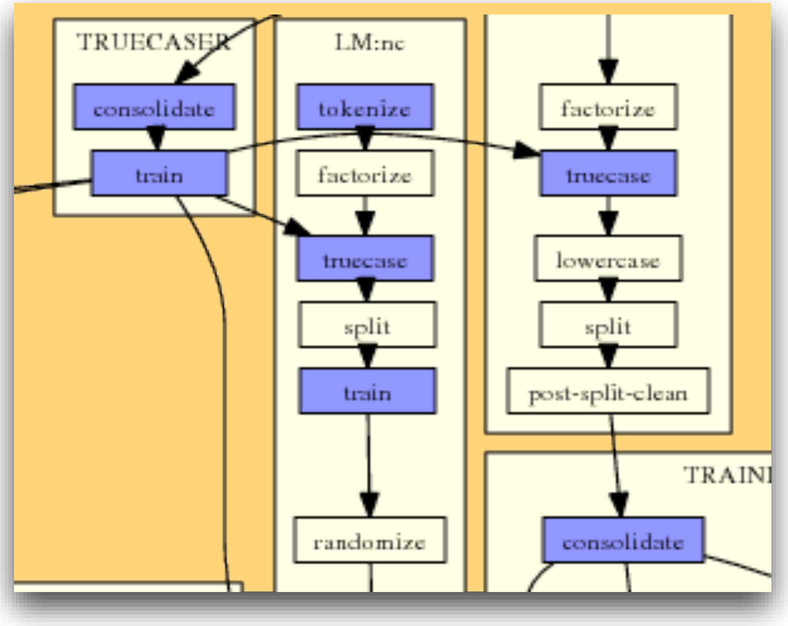
- One configuration file for all settings: record of all experimental details

- Scheduler of individual steps in pipeline
  - automatically keeps track of dependencies
  - on single machine, multi-core machines, GridEngine clusters
  - parallel execution
  - crash detection
  - automatic re-use of prior results

- Fast to use
  - set up a new experiments in minutes
  - set up a variation of an experiment in seconds

Workflow automatically generated by experiment.perl





How does it work?

- Write a configuration file (typically by adapting an existing file)
- Execute: `experiment.perl -config config -exec`

Components

- `experiment.perl`
  - executable that schedules all steps
  - has functions that define more complex steps (e.g., tuning)
- `experiment.meta`
  - meta-configuration file
  - defines all steps and their dependencies
  - template definitions for simpler steps
- `config`
  - includes settings for one experimental run

Jargon

experiment: consists of a number of **runs** that share the same data, same processing **steps**, etc.
example: WMT 2010 German-English system

run: individual experimental instance with specific settings and one outcome
example: run with higher distortion limit

step: processing step of a **run**
examples: tokenization, decoding

module: logical block of processing **steps**
examples: corpus preparation, tuning

set: corpus for training or language modeling
examples: Europarl, News Commentary

setting: specific parameter in the configuration of a experimental **run**
example: `decoder-setting = "-dl 8"`

experiment.meta

- Definition of potential steps of an experiment

- Grouped in modules:
  - CORPUS: preparing a parallel corpus
  - INPUT-FACTOR and OUTPUT-FACTOR: commands to create factors
  - TRAINING: training a translation model
  - LM: training a language model
  - INTERPOLATED-LM: interpolate language models
  - SPLITTER: training a word splitting model
  - RECASING: training a recaser
  - TRUECASING: training a truecaser
  - TUNING: running minumum error rate training to set component weights
  - TESTING: translating and scoring a test set
  - REPORTING: compile all scores in one file

Step Definition

```
[LM]
get-corpus
       in: get-corpus-script
       out: raw-corpus
       [...]

tokenize
       in: raw-corpus
       out: tokenized-corpus
       [...]
```

- Input and outputs establish dependencies between steps (as in a `Makefile`)
  - `tokenize` requires as input `raw-corpus`
  - `get-corpus` produces as output `raw-corpus`
  - when `tokenize` needs to be run, first `raw-corpus` needs to be executed

Configuration File

- Input to a step may be specified in configuration file (`config`):

```
[LM:europarl]

### raw corpus file
#
raw-corpus = $europarl-v3/training/europarl-v3.en
```

- May limit which steps need to be executed
  - `tokenize` requires as input `raw-corpus`
  - `raw-corpus` is specified in the `config` file
  - no need to run `get-corpus`

Elements of Step Definitions

- Several parameters for step definitions are used in `experiment.meta`:
  - `in` and `out`: established dependencies between steps
  - `default-name`: file name of output
  - `template`: template for the command that executes step
  - `pass-unless`: only execute if the given setting is used
  - `error`: if STDERR contains specified key words, step has failed
  - `rerun-on-change`: limits re-use if specified settings are changed
- There are more (see paper or documentation)

Definition of LM:tokenize

```
tokenize
        in: raw-corpus
        out: tokenized-corpus
        default-name: lm/tok
        pass-unless: output-tokenizer
        template: $output-tokenizer < IN > OUT
        parallelizable: yes
```

Configuration File

- List of settings
- Comments and empty lines for better readability
- Organized in sections for each module
  - start of section indicated by module (and set) name
  - examples: `[TRAINING]` or `[CORPUS:europarl]`
- Syntax of setting definition: `setting = value`

Configuration File: Syntax

- Settings can be used as variables to define other settings:

```
working-dir = /home/pkoehn/experiment
wmt10-data = $working-dir/data
```

- Variable names may be placed in curly brackets for clearer separation:

```
wmt10-data = ${working-dir}/data
```

- References to output of other steps

```
[RECASING]
tokenized = [LM:europarl:tokenized-corpus]
```

Step Files

- Command to execute is stored in a file

- After execution, other files are created:

```
steps/1/LM_europarl_tokenize.1
steps/1/LM_europarl_tokenize.1.DONE
steps/1/LM_europarl_tokenize.1.INFO
steps/1/LM_europarl_tokenize.1.STDERR
steps/1/LM_europarl_tokenize.1.STDERR.digest
steps/1/LM_europarl_tokenize.1.STDOUT
```

  - meta information (INFO, DONE)
  - output (STDERR, STDOUT)
  - digest of output for indicators of crash (STDERR.digest)

Re-Use of Steps

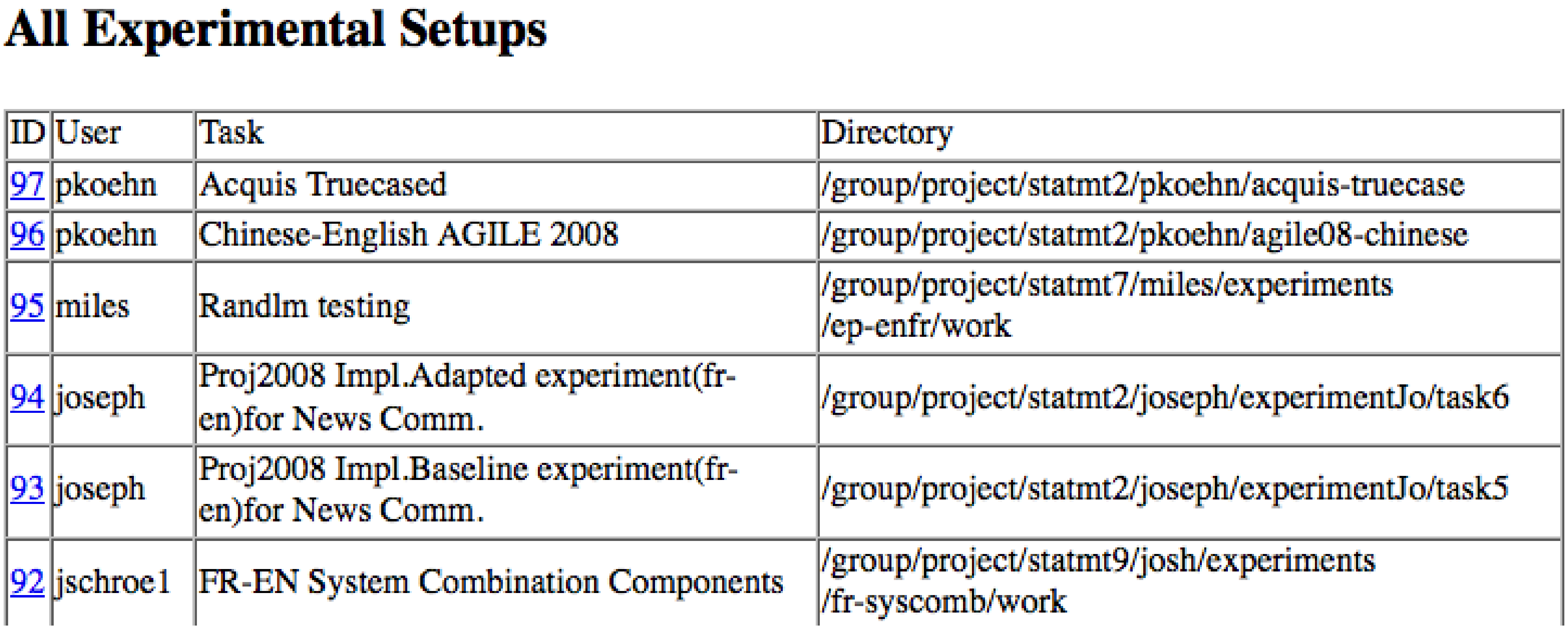
- Example:
  - run 1: baseline
  - run 2: change order of language model

  → tokenization and truecasing of language model training data can be re-used

- Files in directory for language model data:

```
% ls -tr lm/*
lm/europarl.tok.1
lm/europarl.truecased.1
lm/europarl.lm.1
lm/europarl.lm.2
```

Web Interface

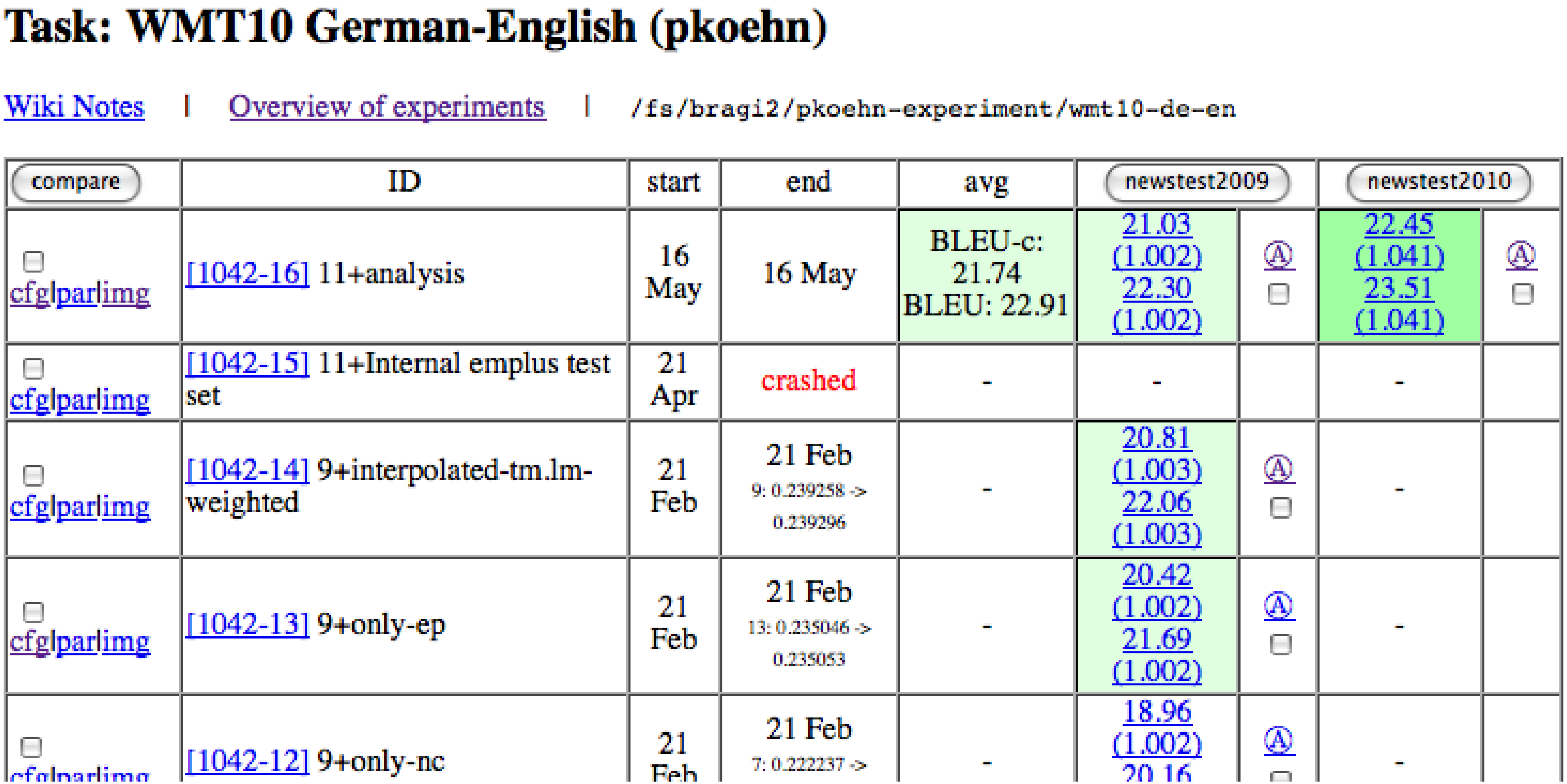


All Experimental Setups

ID	User	Task	Directory
97	pkoehn	Acquis Truecased	/group/project/statmt2/pkoehn/acquis-truecase
96	pkoehn	Chinese-English AGILE 2008	/group/project/statmt2/pkoehn/agile08-chinese
95	miles	Randlm testing	/group/project/statmt7/miles/experiments/ep-enfr/work
94	joseph	Proj2008 Impl.Adapted experiment(fr-en)for News Comm.	/group/project/statmt2/joseph/experimentJo/task6
93	joseph	Proj2008 Impl.Baseline experiment(fr-en)for News Comm.	/group/project/statmt2/joseph/experimentJo/task5
92	jschroe1	FR-EN System Combination Components	/group/project/statmt9/josh/experiments/fr-syscomb/work

List of experiments

List of Runs

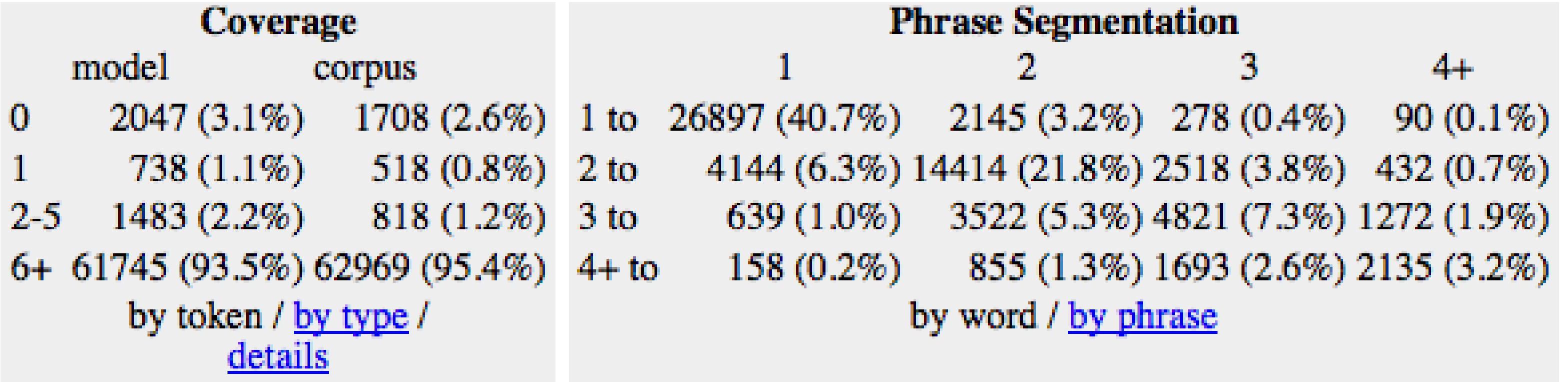


Task: WMT10 German-English (pkoehn)

Wiki Notes | Overview of experiments | /fs/bragi2/pkoehn-experiment/wmt10-de-en

compare	ID	start	end	avg	newstest2009		newstest2010	
□ cfg|par|img	[1042-16] 11+analysis	16 May	16 May	BLEU-c: 21.74 BLEU: 22.91	21.03 (1.002) 22.30 (1.002)	Ⓐ □	22.45 (1.041) 23.51 (1.041)	Ⓐ □
□ cfg|par|img	[1042-15] 11+Internal emplus test set	21 Apr	crashed	-	-		-	
□ cfg|par|img	[1042-14] 9+interpolated-tm.lm-weighted	21 Feb	21 Feb 9: 0.239258 -> 0.239296	-	20.81 (1.003) 22.06 (1.003)	Ⓐ □	-	
□ cfg|par|img	[1042-13] 9+only-ep	21 Feb	21 Feb 13: 0.235046 -> 0.235053	-	20.42 (1.002) 21.69 (1.002)	Ⓐ □	-	
□ cfg|par|img	[1042-12] 9+only-nc	21 Feb	21 Feb 7: 0.222237 ->	-	18.96 (1.002) 20.16	Ⓐ □	-	

Analysis: Basic Statistics



Coverage		
	model	corpus
0	2047 (3.1%)	1708 (2.6%)
1	738 (1.1%)	518 (0.8%)
2-5	1483 (2.2%)	818 (1.2%)
6+	61745 (93.5%)	62969 (95.4%)

by token / by type / details

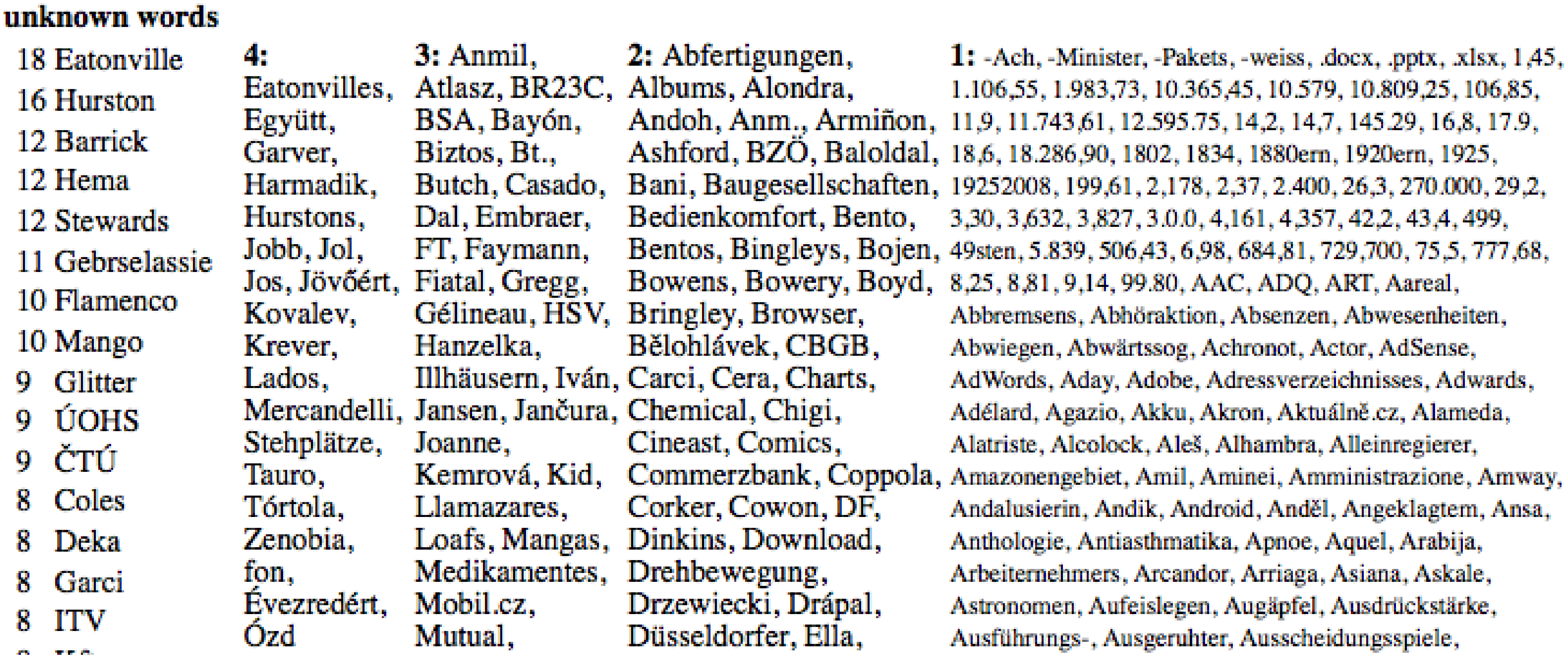
Phrase Segmentation				
	1	2	3	4+
1 to	26897 (40.7%)	2145 (3.2%)	278 (0.4%)	90 (0.1%)
2 to	4144 (6.3%)	14414 (21.8%)	2518 (3.8%)	432 (0.7%)
3 to	639 (1.0%)	3522 (5.3%)	4821 (7.3%)	1272 (1.9%)
4+ to	158 (0.2%)	855 (1.3%)	1693 (2.6%)	2135 (3.2%)

by word / by phrase

- Basic statistics
  - n-gram precision
  - evaluation metrics
  - coverage of the input in corpus and translation model
  - phrase segmentations used

Analysis: Unknown Words

grouped by frequency in test set



unknown words

18	Eatonville
16	Hurston
12	Barrick
12	Hema
12	Stewards
11	Gebrselassie
10	Flamenco
10	Mango
9	Glitter
9	ÚOHS
9	ČTÚ
8	Coles
8	Deka
8	Garci
8	ITV

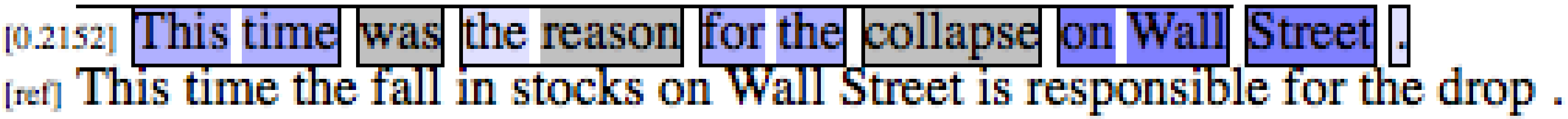
4: Eatonvilles, Együtt, Garver, Harmadik, Hurstons, Jobb, Jol, Jos, Jövőért, Kovalev, Krever, Lados, Mercandelli, Stehplätze, Tauro, Tórtola, Zenobia, fon, Évezredért, Ózd

3: Anmil, Atlasz, BR23C, BSA, Bayón, Biztos, Bt., Butch, Casado, Dal, Embraer, FT, Faymann, Fiatal, Gregg, Gélineau, HSV, Hanzelka, Illhäusern, Iván, Jansen, Jančura, Joanne, Kemrová, Kid, Llamazares, Loafs, Mangas, Medikamentes, Mobil.cz, Mutual,

2: Abfertigungen, Albums, Alondra, Andoh, Anm., Armiñon, Ashford, BZÖ, Baloldal, Bani, Baugesellschaften, Bedienkomfort, Bento, Bentos, Bingleys, Bojen, Bowens, Bowery, Boyd, Bringley, Browser, Bělohlávek, CBGB, Carci, Cera, Charts, Chemical, Chigi, Cineast, Comics, Commerzbank, Coppola, Corker, Cowon, DF, Dinkins, Download, Drehbewegung, Drzewiecki, Drápal, Düsseldorfer, Ella,

1: -Ach, -Minister, -Pakets, -weiss, .docx, .pptx, .xlsx, 1,45, 1.106,55, 1.983,73, 10.365,45, 10.579, 10.809,25, 106,85, 11,9, 11.743,61, 12.595.75, 14,2, 14,7, 145.29, 16,8, 17.9, 18,6, 18.286,90, 1802, 1834, 1880ern, 1920ern, 1925, 19252008, 199,61, 2,178, 2,37, 2.400, 26,3, 270.000, 29,2, 3,30, 3,632, 3,827, 3.0.0, 4,161, 4,357, 42,2, 43,4, 499, 49sten, 5.839, 506,43, 6,98, 684,81, 729,700, 75,5, 777,68, 8,25, 8,81, 9,14, 99.80, AAC, ADQ, ART, Aareal, Abbremsens, Abhöraktion, Absenzen, Abwesenheiten, Abwiegen, Abwärtssog, Achronot, Actor, AdSense, AdWords, Aday, Adobe, Adressverzeichnisses, Adwards, Adélard, Agazio, Akku, Akron, Aktuálně.cz, Alameda, Alatriste, Alcolock, Aleš, Alhambra, Alleinregierer, Amazonengebiet, Amil, Aminei, Amministrazione, Amway, Andalusierin, Andik, Android, Anděl, Angeklagtem, Ansa, Anthologie, Antiasthmatika, Apnoe, Aquel, Arabija, Arbeiternehmers, Arcandor, Arriaga, Asiana, Askale, Astronomen, Aufeislegen, Augäpfel, Ausdrückstärke, Ausführungs-, Ausgeruhter, Ausscheidungsspiele,

Analysis: Output Annotation



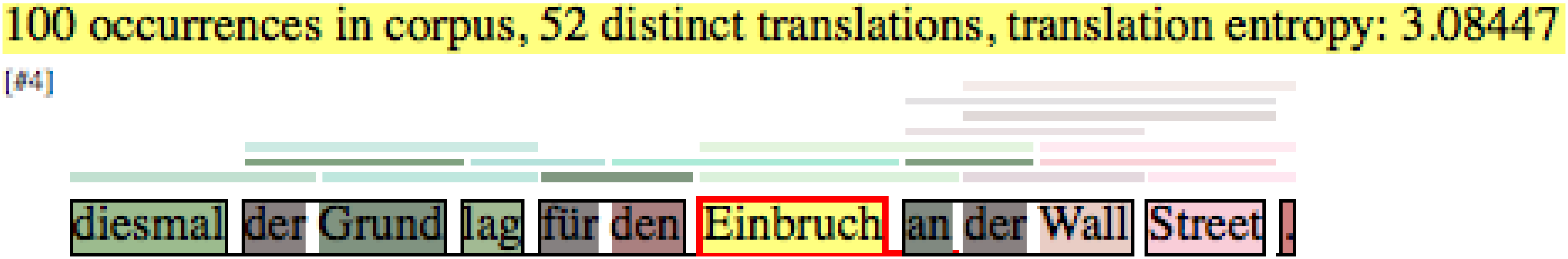
[0.2152] This time was the reason for the collapse on Wall Street .

[ref] This time the fall in stocks on Wall Street is responsible for the drop .

Color highlighting to indicate n-gram overlap with reference translation

darker bleu = word is part of larger n-gram match

Analysis: Input Annotation



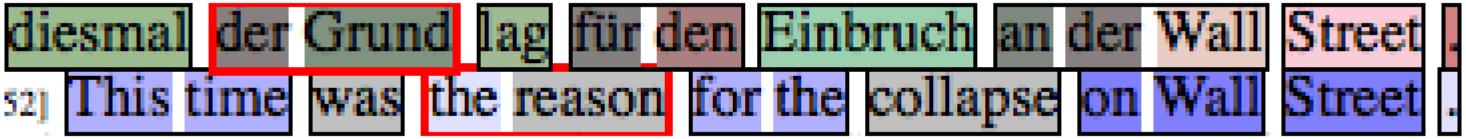
100 occurrences in corpus, 52 distinct translations, translation entropy: 3.08447

[#4]

diesmal der Grund lag für den Einbruch an der Wall Street .

- For each word and phrase, color coding and stats on
  - number of occurrences in training corpus
  - number of distinct translations in translation model
  - entropy of conditional translation probability distribution $\phi(e|f)$ (normalized)

Analysis: Alignment



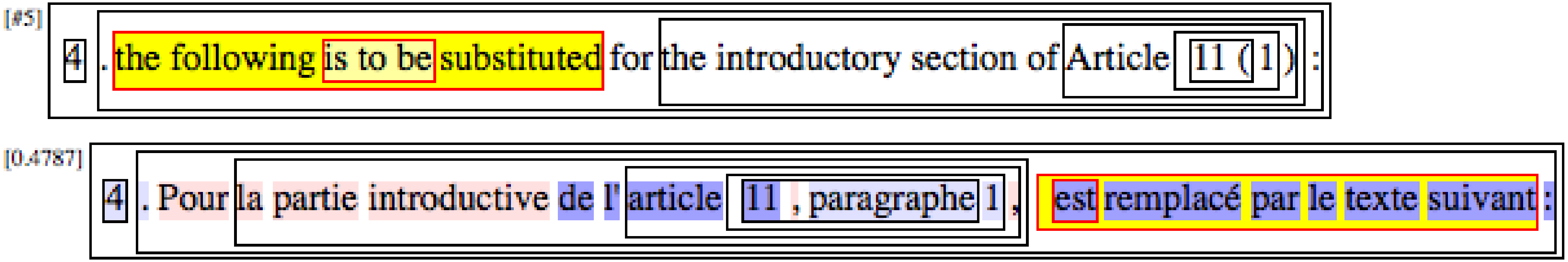
diesmal der Grund lag für den Einbruch an der Wall Street .

52] This time was the reason for the collapse on Wall Street .

Phrase alignment of the decoding process

(red border, interactive)

Analysis: Tree Alignment



[#5]

4 . the following is to be substituted for the introductory section of Article 11 (1) :

[0.4787]

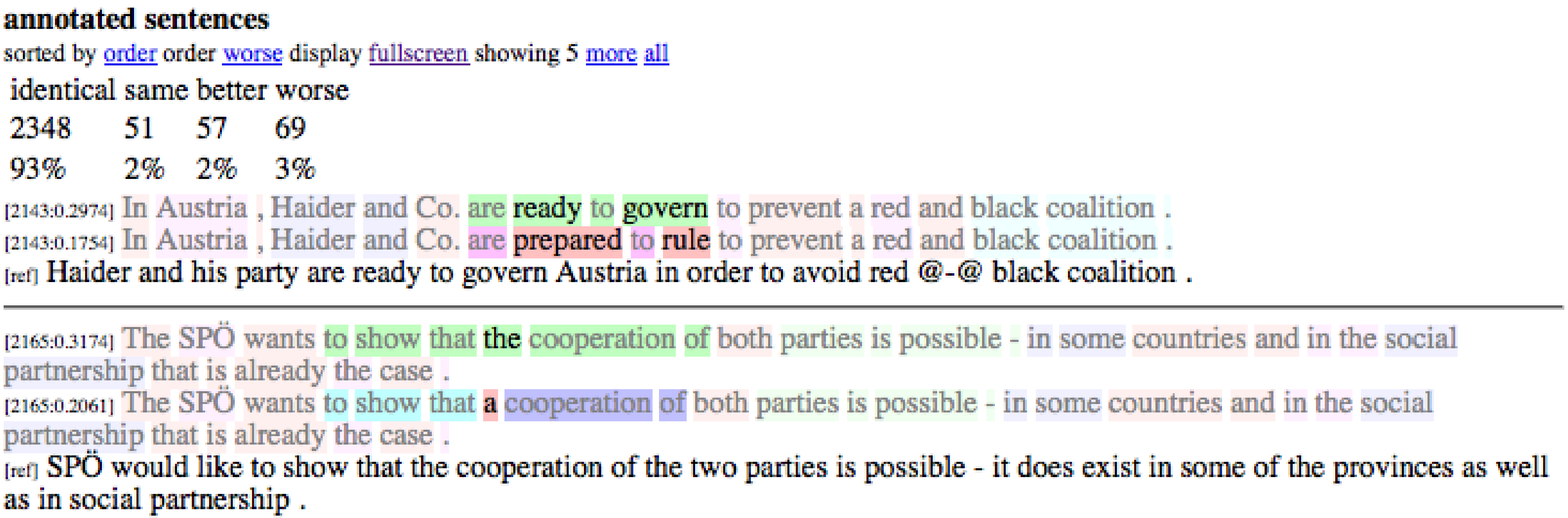
4 . Pour la partie introductive de l' article 11 , paragraphe 1 , est remplacé par le texte suivant :

Uses nested boxes to indicate tree structure

(red border, yellow shaded spans in focus, interactive)

for syntax model, non-terminals are also shown

Analysis: Comparison of 2 Runs



annotated sentences

sorted by order order worse display fullscreen showing 5 more all

identical	same	better	worse
2348	51	57	69
93%	2%	2%	3%

[2143:0.2974] In Austria , Haider and Co. are ready to govern to prevent a red and black coalition .
[2143:0.1754] In Austria , Haider and Co. are prepared to rule to prevent a red and black coalition .
[ref] Haider and his party are ready to govern Austria in order to avoid red @-@ black coalition .

[2165:0.3174] The SPÖ wants to show that the cooperation of both parties is possible - in some countries and in the social partnership that is already the case .
[2165:0.2061] The SPÖ wants to show that a cooperation of both parties is possible - in some countries and in the social partnership that is already the case .
[ref] SPÖ would like to show that the cooperation of the two parties is possible - it does exist in some of the provinces as well as in social partnership .

Different words are highlighted

sortable by most improvement, deterioration

Conclusion

- Experiment.perl makes life easier
  - setting up complex experiments with one configuration file
  - permanent record of parameter settings
  - easily distributed (Edinburgh's WMT 2010 system configs available)

- Analysis allows insight into model performance
  - basic stats
  - inspect derivations and options of decoder
  - differences between two runs

- Future plans
  - integrate more tools (also yours, help wanted!)
  - scheduling jobs on Hadoop
  - more analysis