Towards Robust Semantic Role Labeling

Pradhan, Ward and Martin (2008)

Seminar: Recent Developments in Computational Semantics

Ruth Schreiber
June 21st 2010

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

What is SRL?

- annotate naturally occurring text with semantic structure; used for
  - information extraction
  - question answering
  - summarization
- for each predicate in a sentence, identify and label its semantic arguments

  [AGENT John] **broke** [THEME the window]

State of the Art

- supervised machine learning with hand-corrected syntactic parses
  - good accuracies
- recent approaches: improved features, e.g. n-best parses
- good performances on standard test data, but: performance decreases significantly when test data are drawn from a genre different from the data on which the system was trained
  - WHY?

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

Semantic Annotation and Corpora

- reproduce the semantic labeling scheme used by the PropBank corpus
  - predicate independent labels (core arguments ARG0 to ARG5, adjunctive arguments ARGMS)
- assumption: semantic argument of a predicate aligns with one or more nodes in the hand-corrected Treebank parses

Semantic Annotation and Corpora (2)

- Corpus: Wall Street Journal (WSJ)
- training set: Section 02 to Section 21 (~ 90,000 predicates)
- development set: Section 00
- test set: Section 23 (~ 5,000 predicates)

Task Description

- assign role labels to constituents of a syntactic parse
- three different tasks

  (1) Argument Identification: classify NON-NULL and NULL nodes

  (2) Argument Classification: assign appropriate argument labels to given constituents known to represents an argument

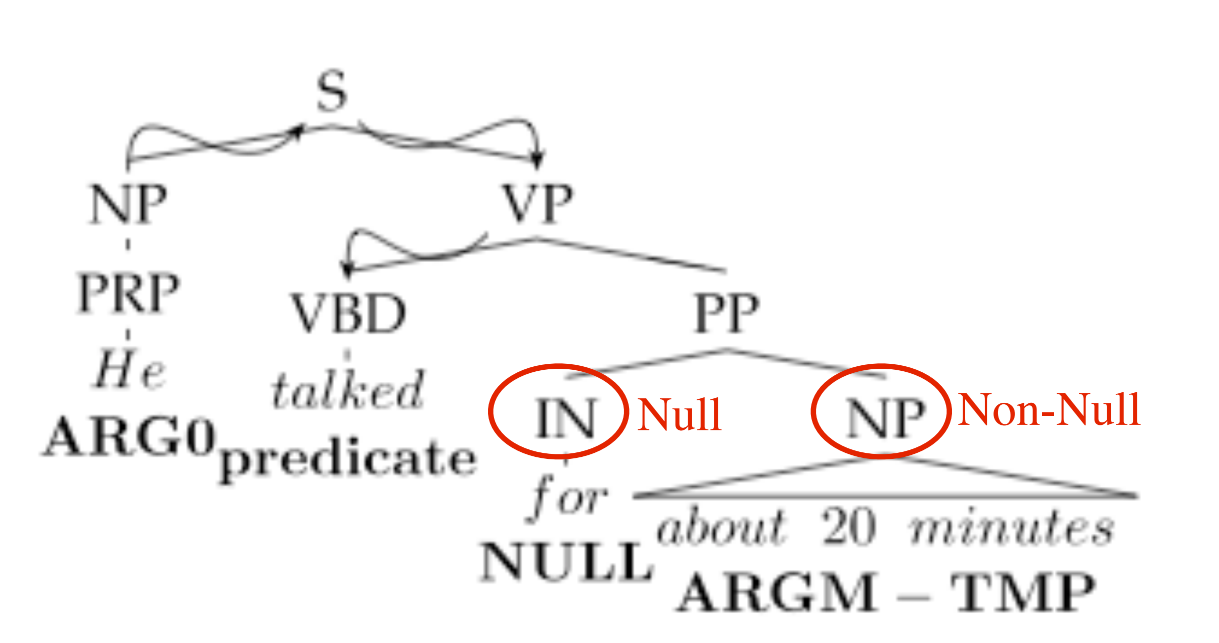
  (3) Argument Identification & Classification

Task Description

- assign role labels to constituents of a syntactic parse
- three different tasks

(1)Argument Identification: classify NON-NULL and NULL nodes

(2)Arg gn app given con ts an argu



(3)Arg assification

Task Description

- assign role labels to constituents of a syntactic parse
- three different tasks

(1) Argument Identification: classify NON-NULL and NULL nodes

(2) Argument Classification: assign appropriate argument labels to given constituents known to represents an argument

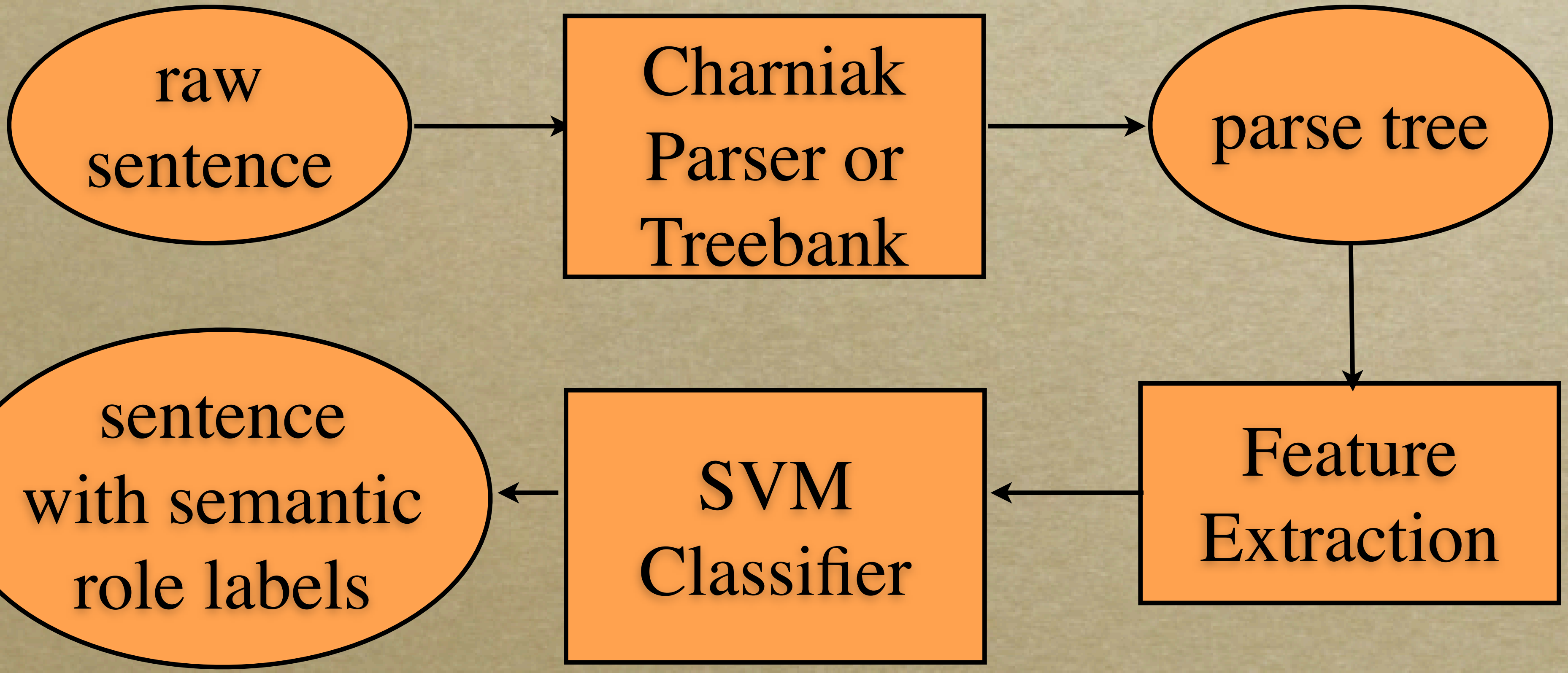
(3) Argument Identification & Classification

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

ASSERT: Architecture

- produces semantic role labels (PropBank arguments and NULL) for each non-copula/non-auxiliary predicate in a sentence


raw sentence
Charniak Parser or Treebank
parse tree
Feature Extraction
SVM Classifier
sentence with semantic role labels

ASSERT: Architecture (2)

- multi-class classification problem using SVMs
  - ONE vs. ALL approach
- drawback: each argument classification is made independently
  - ignores that each predicate is likely to instantiate a certain set of arguments
  - solution: train backed-off trigram model for argument sequences

ASSERT: Architecture (3)

- system generates argument lattice using n-best hypotheses for each node in the syntax tree
- Viterbi search through lattice: uses observation probabilities and language model probabilities
- goal: find maximum likelihood path such that each node is assigned a label (PropBank or NULL)
- no two overlapping nodes are both assigned NON−NULL lables

ASSERT: Features

- Predicate: surface form and lemma
- Path: syntactic parse through parse tree from constituent to predicate
- Phrase Type: syntactic category of constituent
- Partial Path: Path from constituent to lowest common ancestor of predicate and constituent
- Head Word: Syntactic head of constituent
- First and Last Word/POS in constituent

ASSERT: Performance

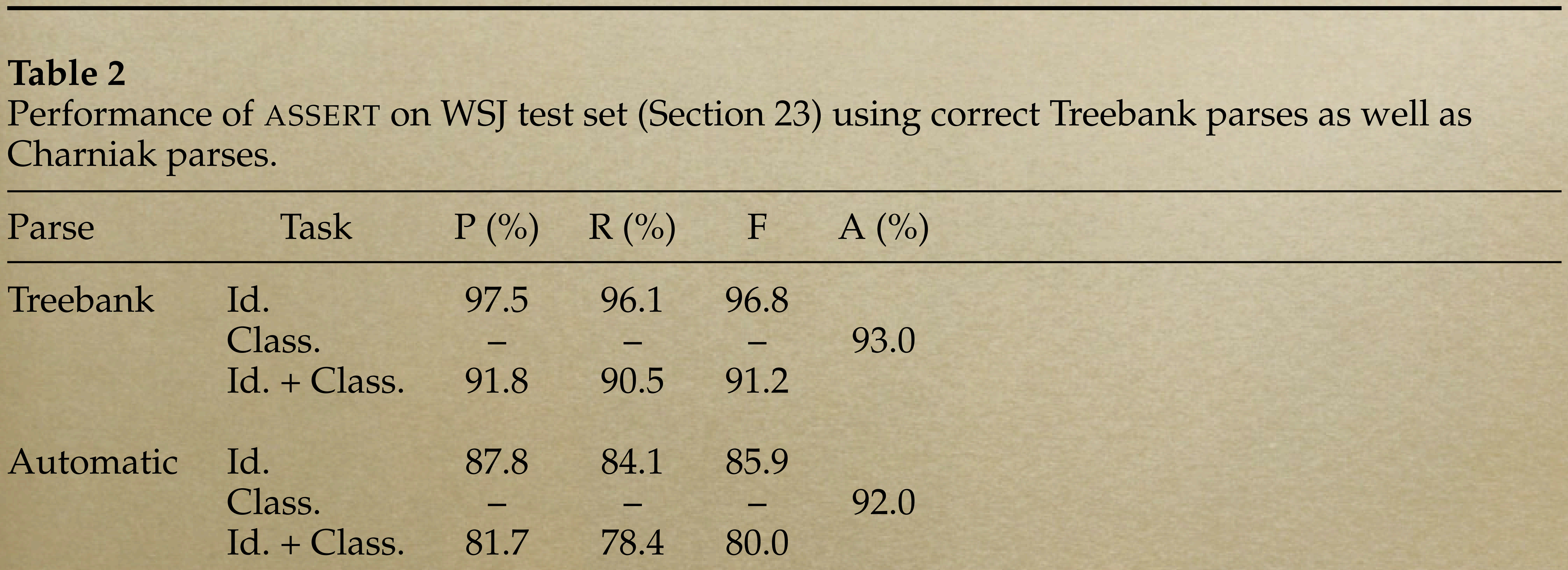


Table 2
Performance of ASSERT on WSJ test set (Section 23) using correct Treebank parses as well as Charniak parses.

Parse	Task	P (%)	R (%)	F	A (%)
Treebank	Id.	97.5	96.1	96.8	
	Class.	–	–	–	93.0
	Id. + Class.	91.8	90.5	91.2	
Automatic	Id.	87.8	84.1	85.9	
	Class.	–	–	–	92.0
	Id. + Class.	81.7	78.4	80.0	

ASSERT: Performance

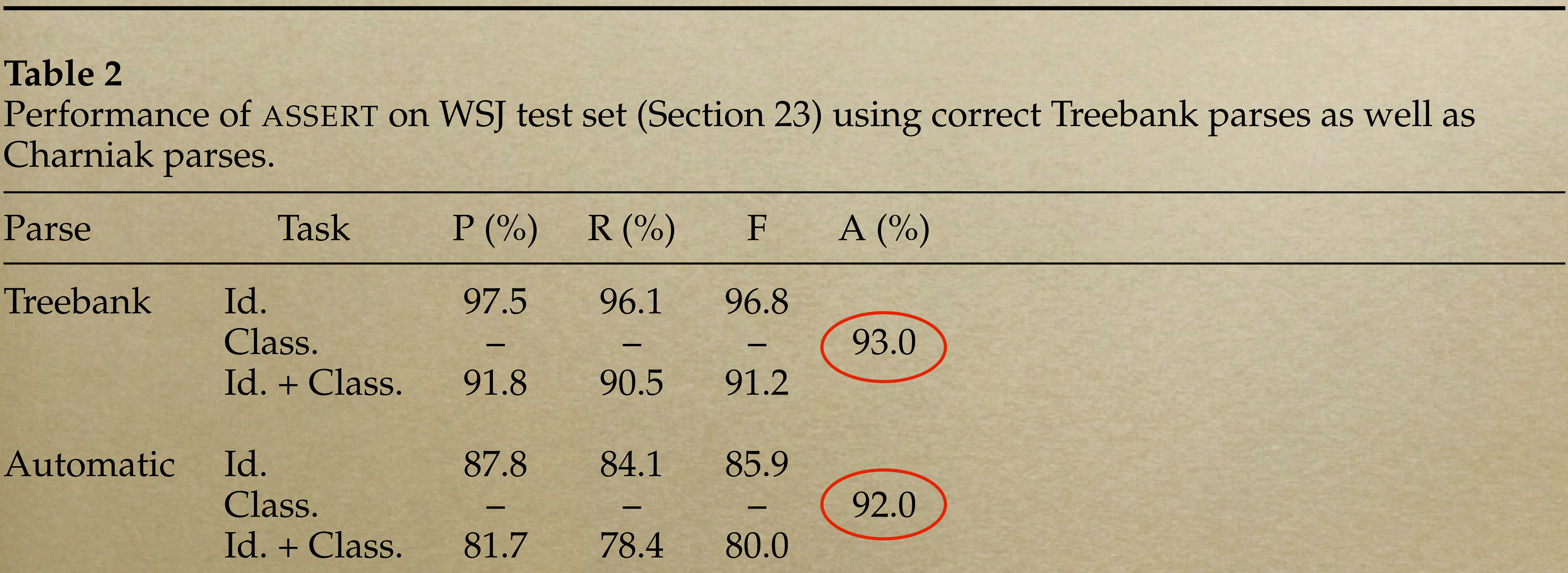


Table 2
Performance of ASSERT on WSJ test set (Section 23) using correct Treebank parses as well as Charniak parses.

Parse	Task	P (%)	R (%)	F	A (%)
Treebank	Id.	97.5	96.1	96.8	
	Class.	–	–	–	93.0
	Id. + Class.	91.8	90.5	91.2	
Automatic	Id.	87.8	84.1	85.9	
	Class.	–	–	–	92.0
	Id. + Class.	81.7	78.4	80.0	

ASSERT: Performance

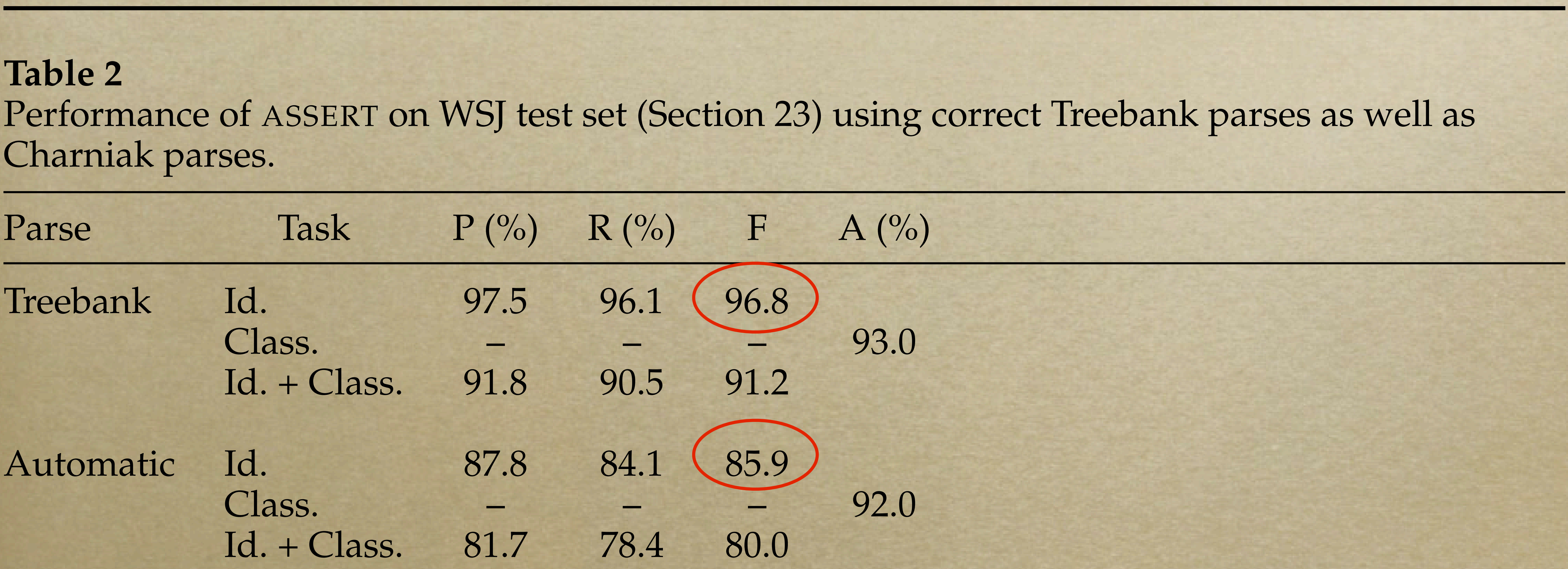


Table 2
Performance of ASSERT on WSJ test set (Section 23) using correct Treebank parses as well as Charniak parses.

Parse	Task	P (%)	R (%)	F	A (%)
Treebank	Id.	97.5	96.1	96.8	
	Class.	–	–	–	93.0
	Id. + Class.	91.8	90.5	91.2	
Automatic	Id.	87.8	84.1	85.9	
	Class.	–	–	–	92.0
	Id. + Class.	81.7	78.4	80.0	

ASSERT: Feature Salience

- structural features (like Path and Partial Path) are important for Identification task
- lexical/semantic features (like Predicate, Head Word, First Word, Last Word) are important for Classification task

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

Robustness to Genre of Data

- most SRL systems have only been tested on a test set belonging to the same genre of text as the training set
  - improvements to the systems could be due to tuning to the specific data
- important that SRL systems also perform well on texts from a different genre than the training data
  - evaluate performance of ASSERT on test data from Brown Corpus and understand which factors affect performance

The Brown Corpus

- standard corpus of American English (about one million words)
- 15 different sections, e.g. General Fiction, Science Fiction, Adventure, Humor
- Penn Treebank contains parses for a subset of the Brown Corpus
- PropBank roles available for parts of the Treebanked Brown Corpus

Cross-Genre Testing

- ASSERT system was trained on WSJ sections 02-21 and evaluated on PropBanked portion of Brown corpus
- parse trees generated by Charniak

Cross-Genre Performance

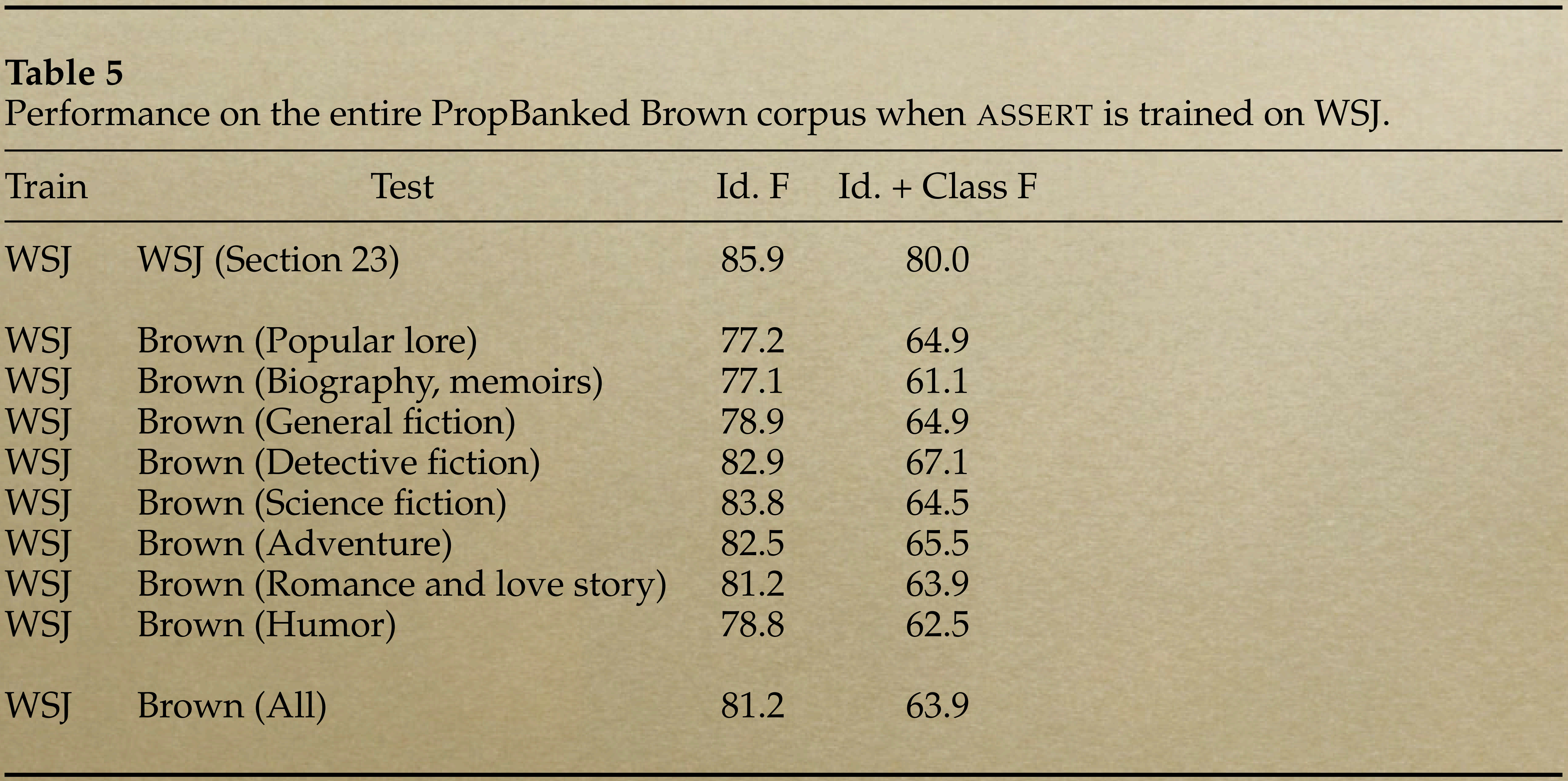


Table 5
Performance on the entire PropBanked Brown corpus when ASSERT is trained on WSJ.

Train	Test	Id. F	Id. + Class F
WSJ	WSJ (Section 23)	85.9	80.0
WSJ	Brown (Popular lore)	77.2	64.9
WSJ	Brown (Biography, memoirs)	77.1	61.1
WSJ	Brown (General fiction)	78.9	64.9
WSJ	Brown (Detective fiction)	82.9	67.1
WSJ	Brown (Science fiction)	83.8	64.5
WSJ	Brown (Adventure)	82.5	65.5
WSJ	Brown (Romance and love story)	81.2	63.9
WSJ	Brown (Humor)	78.8	62.5
WSJ	Brown (All)	81.2	63.9

Cross-Genre Performance

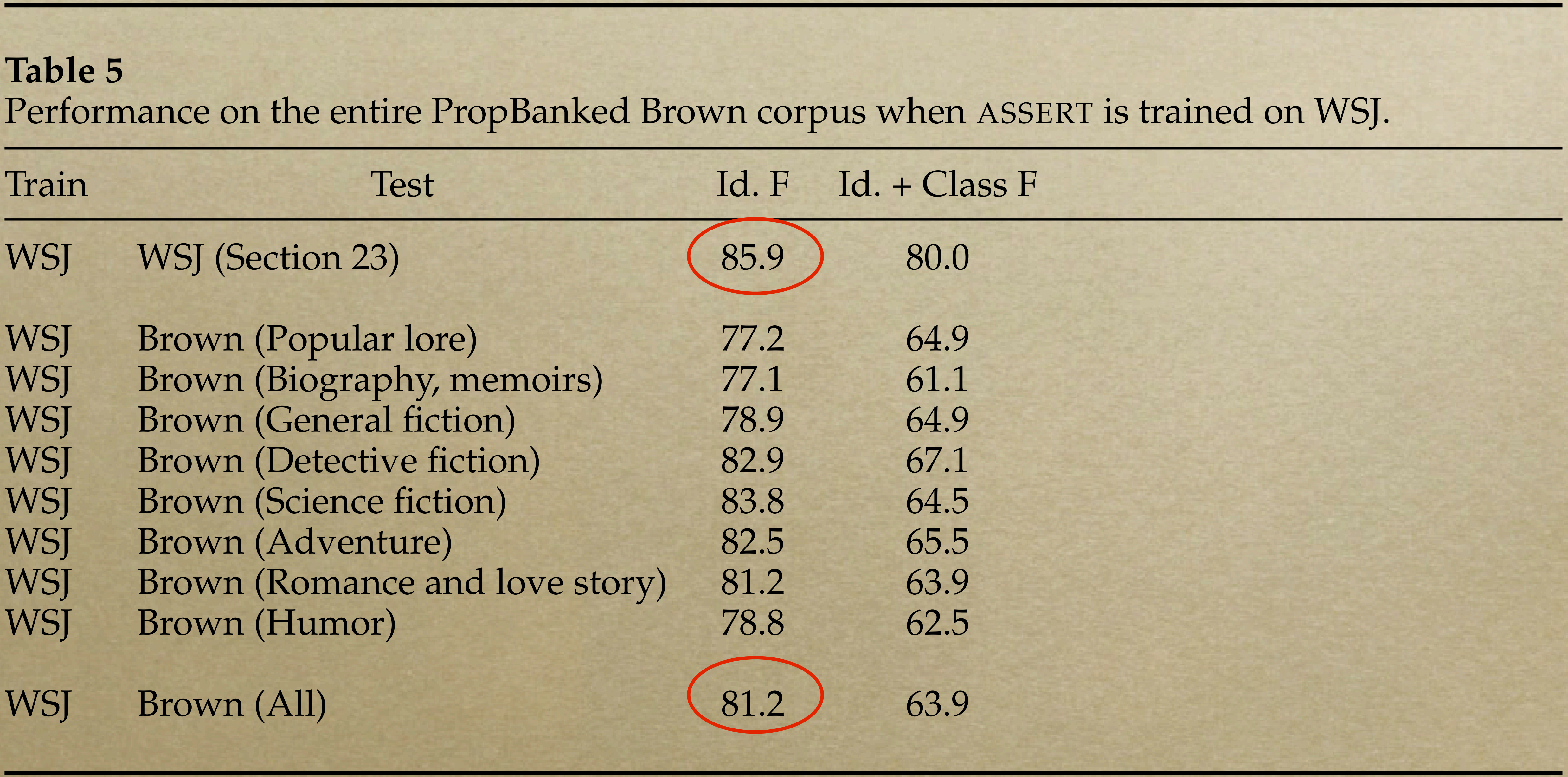


Table 5
Performance on the entire PropBanked Brown corpus when ASSERT is trained on WSJ.

Train	Test	Id. F	Id. + Class F
WSJ	WSJ (Section 23)	85.9	80.0
WSJ	Brown (Popular lore)	77.2	64.9
WSJ	Brown (Biography, memoirs)	77.1	61.1
WSJ	Brown (General fiction)	78.9	64.9
WSJ	Brown (Detective fiction)	82.9	67.1
WSJ	Brown (Science fiction)	83.8	64.5
WSJ	Brown (Adventure)	82.5	65.5
WSJ	Brown (Romance and love story)	81.2	63.9
WSJ	Brown (Humor)	78.8	62.5
WSJ	Brown (All)	81.2	63.9

Cross-Genre Performance

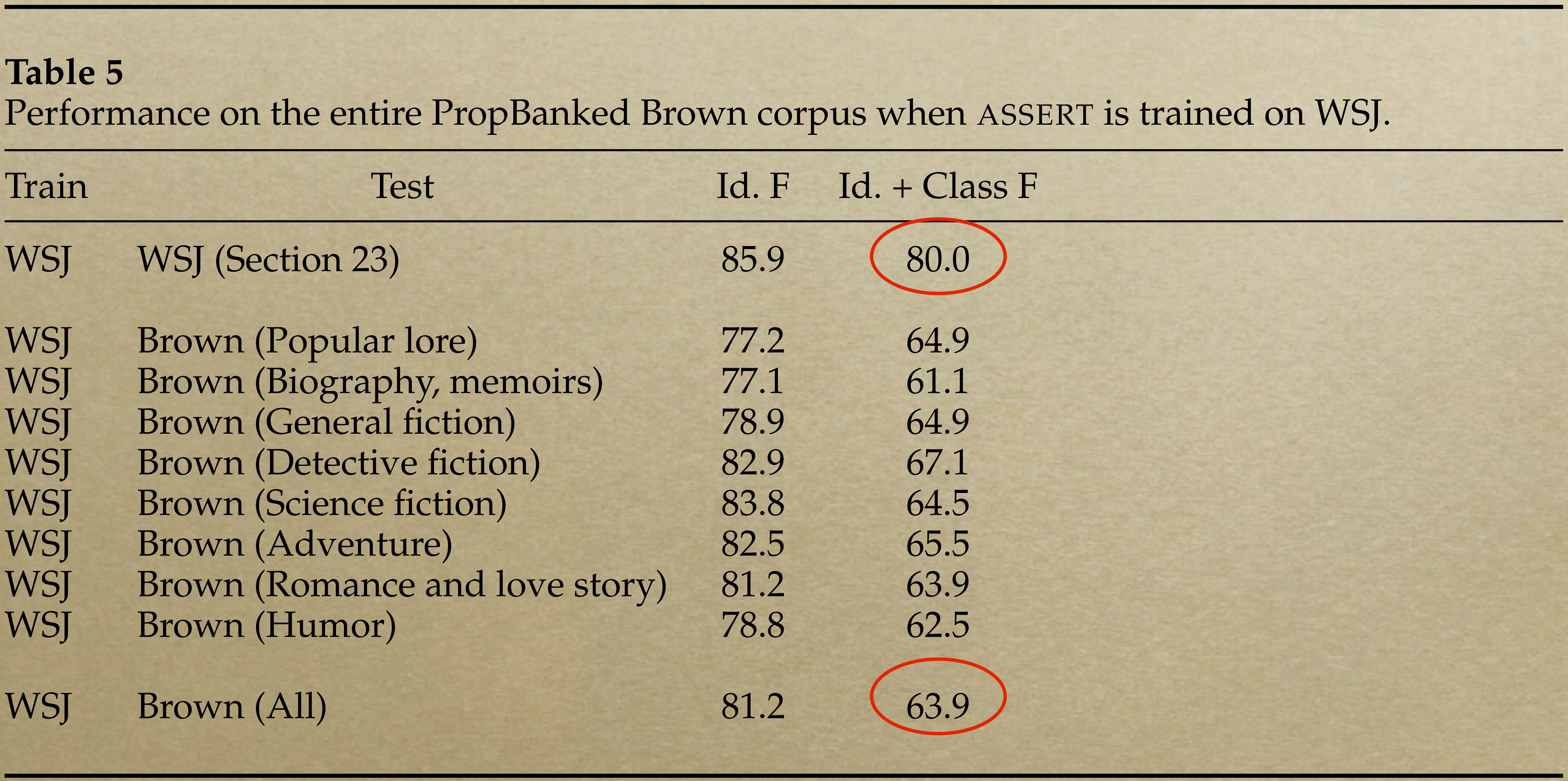


Table 5
Performance on the entire PropBanked Brown corpus when ASSERT is trained on WSJ.

Train	Test	Id. F	Id. + Class F
WSJ	WSJ (Section 23)	85.9	80.0
WSJ	Brown (Popular lore)	77.2	64.9
WSJ	Brown (Biography, memoirs)	77.1	61.1
WSJ	Brown (General fiction)	78.9	64.9
WSJ	Brown (Detective fiction)	82.9	67.1
WSJ	Brown (Science fiction)	83.8	64.5
WSJ	Brown (Adventure)	82.5	65.5
WSJ	Brown (Romance and love story)	81.2	63.9
WSJ	Brown (Humor)	78.8	62.5
WSJ	Brown (All)	81.2	63.9

Possible Reasons for Performance Degradation

- syntactic parsing errors
  - Charniak Parser is heavily lexicalized => genre difference will have an effect on the accuracy of the parses and on the features extracted from them
- difficulty of the corpus
  - greater diversity in use of predicates and headwords => lexical features may be more varied in terms of predicate senses and raw number of predicates

Syntactic Parsing Errors

- argument bearing nodes deleted from the syntactic parse leading to an Identification error: 6.2% for WSJ test set and 8.1% for Brown test set
  - can explain decrease in Identification performance, but not in combined task performance
- effect of errors from syntactic parse can be removed by using the correct syntactic trees from the Treebank for both corpora

Cross-Reference Performance using Treebank parses

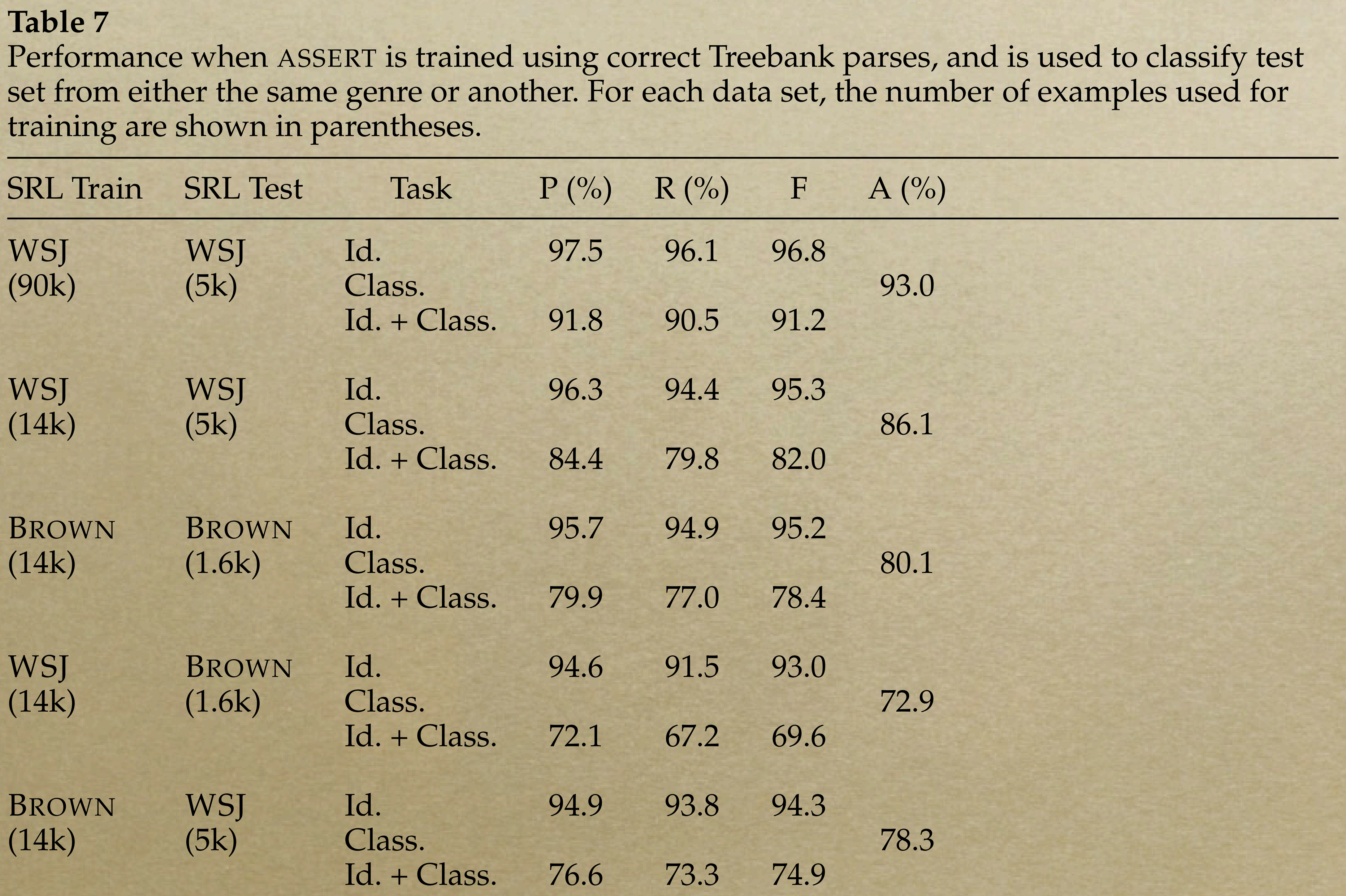


Table 7
Performance when ASSERT is trained using correct Treebank parses, and is used to classify test set from either the same genre or another. For each data set, the number of examples used for training are shown in parentheses.

SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
WSJ	WSJ	Id.	97.5	96.1	96.8	
(90k)	(5k)	Class.				93.0
		Id. + Class.	91.8	90.5	91.2	
WSJ	WSJ	Id.	96.3	94.4	95.3	
(14k)	(5k)	Class.				86.1
		Id. + Class.	84.4	79.8	82.0	
BROWN	BROWN	Id.	95.7	94.9	95.2	
(14k)	(1.6k)	Class.				80.1
		Id. + Class.	79.9	77.0	78.4	
WSJ	BROWN	Id.	94.6	91.5	93.0	
(14k)	(1.6k)	Class.				72.9
		Id. + Class.	72.1	67.2	69.6	
BROWN	WSJ	Id.	94.9	93.8	94.3	
(14k)	(5k)	Class.				78.3
		Id. + Class.	76.6	73.3	74.9	

Cross-Reference Performance using Treebank parses

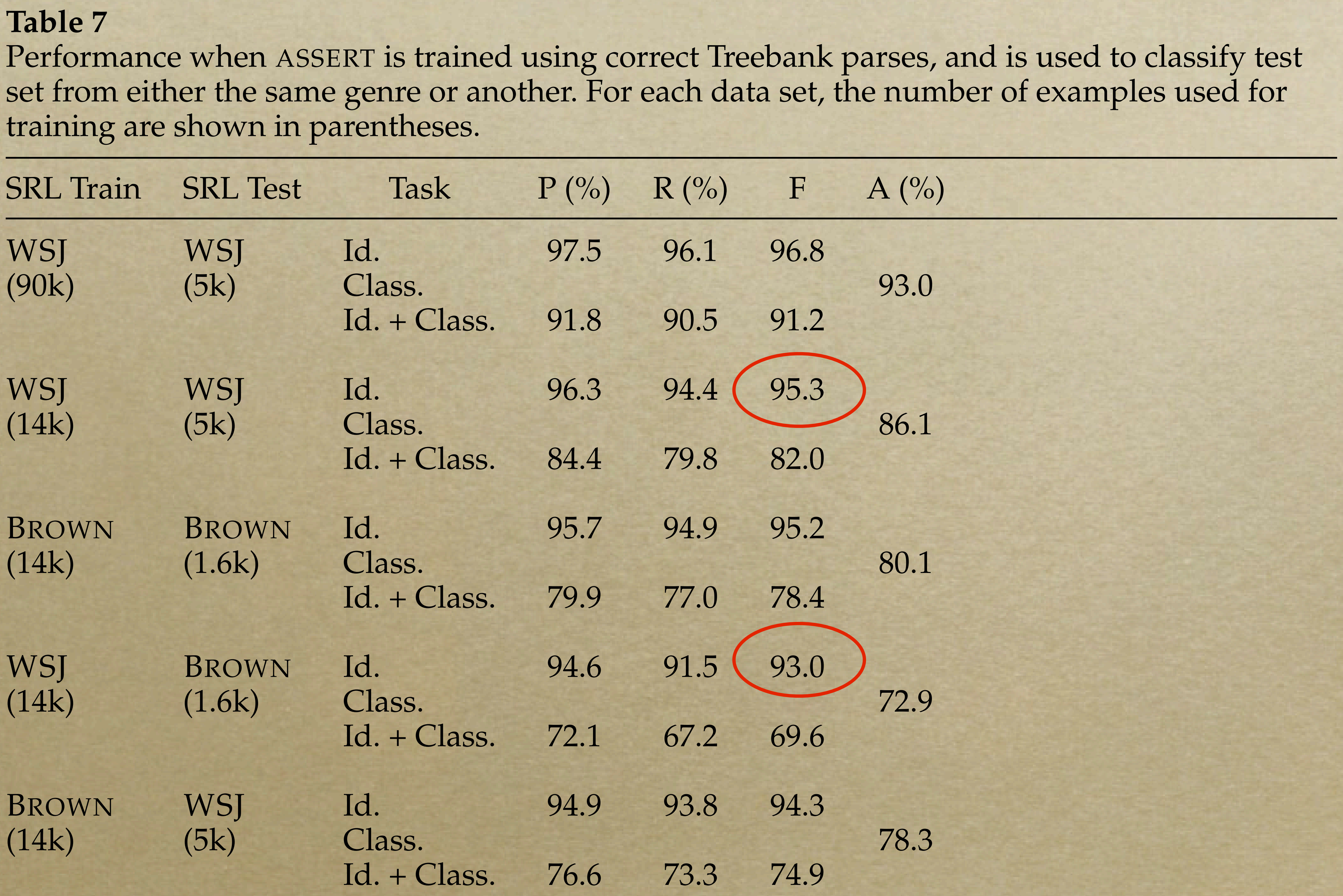


Table 7
Performance when ASSERT is trained using correct Treebank parses, and is used to classify test set from either the same genre or another. For each data set, the number of examples used for training are shown in parentheses.

SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
WSJ	WSJ	Id.	97.5	96.1	96.8	
(90k)	(5k)	Class.				93.0
		Id. + Class.	91.8	90.5	91.2	
WSJ	WSJ	Id.	96.3	94.4	95.3	
(14k)	(5k)	Class.				86.1
		Id. + Class.	84.4	79.8	82.0	
BROWN	BROWN	Id.	95.7	94.9	95.2	
(14k)	(1.6k)	Class.				80.1
		Id. + Class.	79.9	77.0	78.4	
WSJ	BROWN	Id.	94.6	91.5	93.0	
(14k)	(1.6k)	Class.				72.9
		Id. + Class.	72.1	67.2	69.6	
BROWN	WSJ	Id.	94.9	93.8	94.3	
(14k)	(5k)	Class.				78.3
		Id. + Class.	76.6	73.3	74.9	

Cross-Reference Performance using Treebank parses

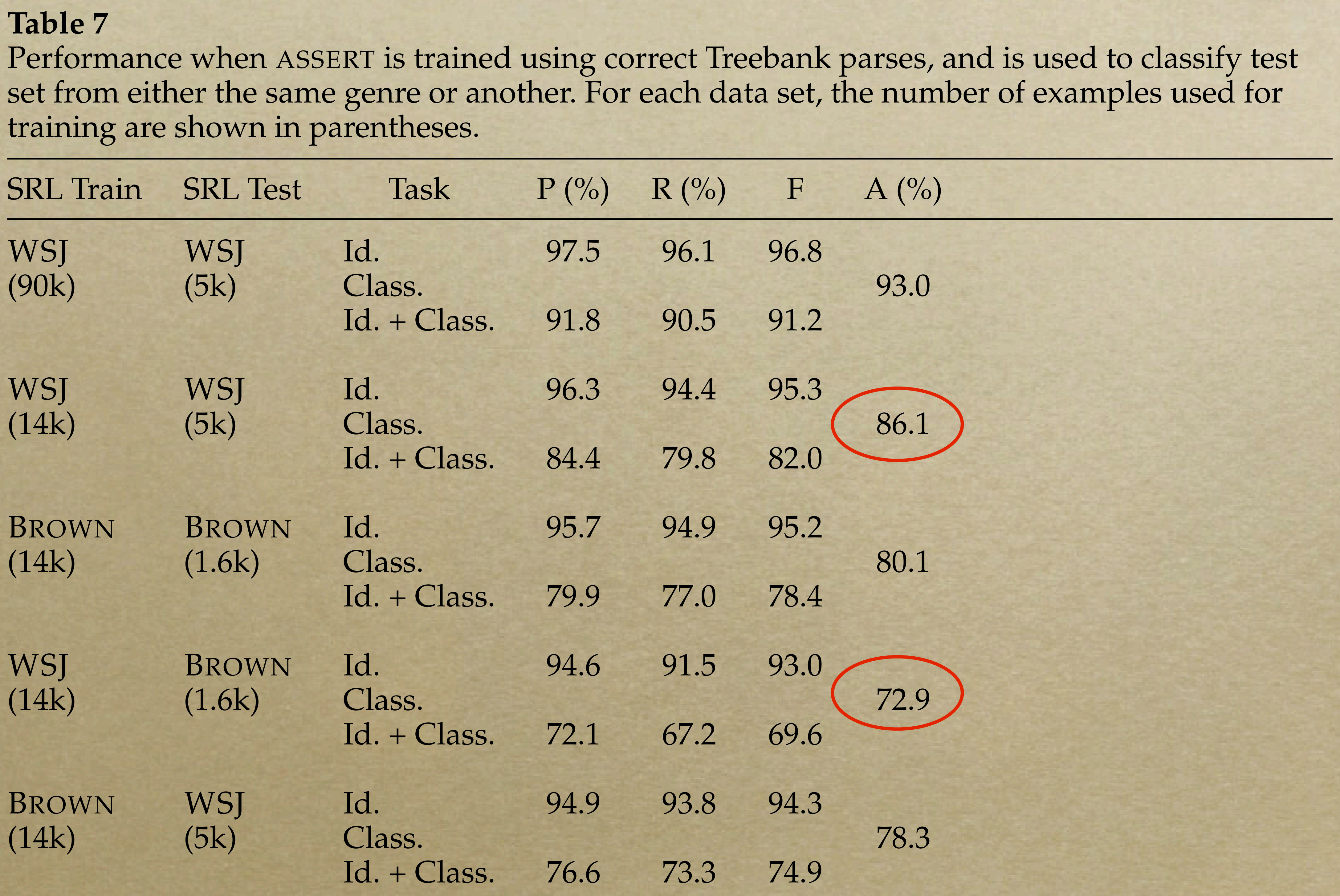


Table 7
Performance when ASSERT is trained using correct Treebank parses, and is used to classify test set from either the same genre or another. For each data set, the number of examples used for training are shown in parentheses.

SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
WSJ	WSJ	Id.	97.5	96.1	96.8	
(90k)	(5k)	Class.				93.0
		Id. + Class.	91.8	90.5	91.2	
WSJ	WSJ	Id.	96.3	94.4	95.3	
(14k)	(5k)	Class.				86.1
		Id. + Class.	84.4	79.8	82.0	
BROWN	BROWN	Id.	95.7	94.9	95.2	
(14k)	(1.6k)	Class.				80.1
		Id. + Class.	79.9	77.0	78.4	
WSJ	BROWN	Id.	94.6	91.5	93.0	
(14k)	(1.6k)	Class.				72.9
		Id. + Class.	72.1	67.2	69.6	
BROWN	WSJ	Id.	94.9	93.8	94.3	
(14k)	(5k)	Class.				78.3
		Id. + Class.	76.6	73.3	74.9	

Cross-Reference Performance using Treebank parses

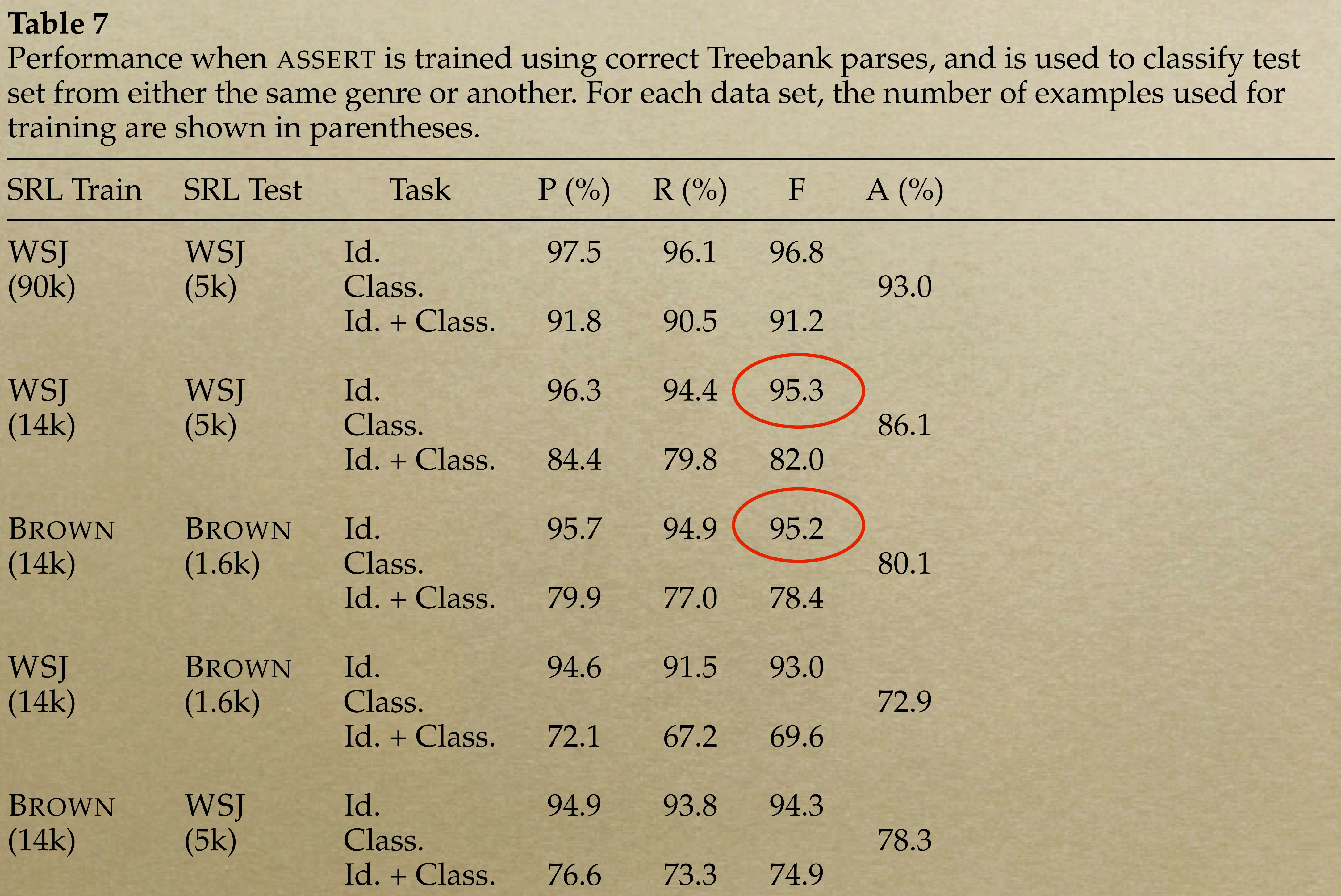


Table 7
Performance when ASSERT is trained using correct Treebank parses, and is used to classify test set from either the same genre or another. For each data set, the number of examples used for training are shown in parentheses.

SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
WSJ	WSJ	Id.	97.5	96.1	96.8	
(90k)	(5k)	Class.				93.0
		Id. + Class.	91.8	90.5	91.2	
WSJ	WSJ	Id.	96.3	94.4	95.3	
(14k)	(5k)	Class.				86.1
		Id. + Class.	84.4	79.8	82.0	
BROWN	BROWN	Id.	95.7	94.9	95.2	
(14k)	(1.6k)	Class.				80.1
		Id. + Class.	79.9	77.0	78.4	
WSJ	BROWN	Id.	94.6	91.5	93.0	
(14k)	(1.6k)	Class.				72.9
		Id. + Class.	72.1	67.2	69.6	
BROWN	WSJ	Id.	94.9	93.8	94.3	
(14k)	(5k)	Class.				78.3
		Id. + Class.	76.6	73.3	74.9	

Cross-Reference Performance using Treebank parses

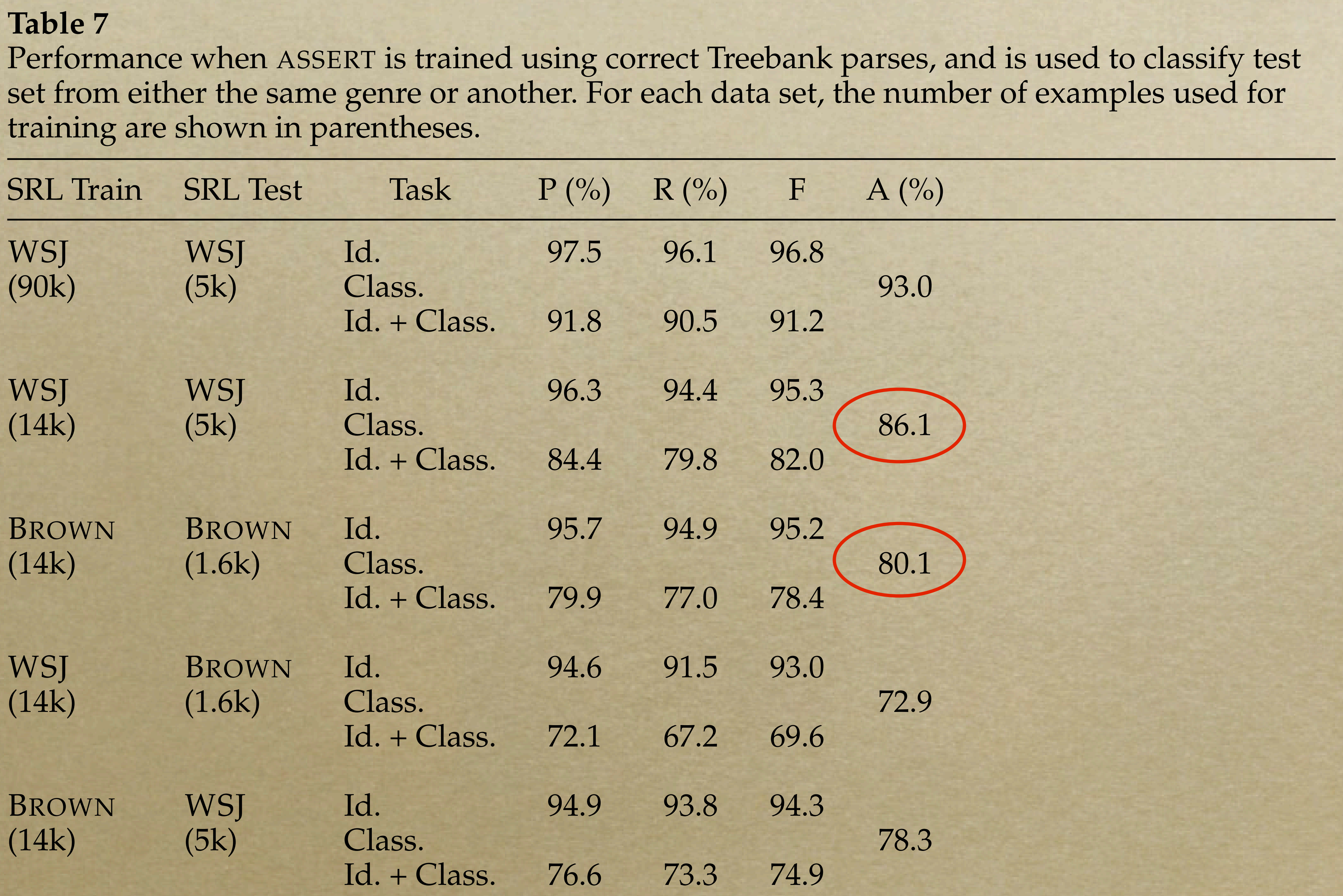


Table 7
Performance when ASSERT is trained using correct Treebank parses, and is used to classify test set from either the same genre or another. For each data set, the number of examples used for training are shown in parentheses.

SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
WSJ	WSJ	Id.	97.5	96.1	96.8	
(90k)	(5k)	Class.				93.0
		Id. + Class.	91.8	90.5	91.2	
WSJ	WSJ	Id.	96.3	94.4	95.3	
(14k)	(5k)	Class.				86.1
		Id. + Class.	84.4	79.8	82.0	
BROWN	BROWN	Id.	95.7	94.9	95.2	
(14k)	(1.6k)	Class.				80.1
		Id. + Class.	79.9	77.0	78.4	
WSJ	BROWN	Id.	94.6	91.5	93.0	
(14k)	(1.6k)	Class.				72.9
		Id. + Class.	72.1	67.2	69.6	
BROWN	WSJ	Id.	94.9	93.8	94.3	
(14k)	(5k)	Class.				78.3
		Id. + Class.	76.6	73.3	74.9	

Difficulty of the Corpus

- more unique predicates or head words than are seen in the WSJ set => less training data for each
- more predicate sense ambiguity in Brown
- less consistent relations between predicates and head words
- more difficult semantic roles in Brown
- relatively fewer examples of predictive features such as named entities

Test Importance of Predicate Sense

- added oracle predicate sense information as a feature in ASSERT
  - 60% of PropBanked Brown tagged with predicate sense information

Test Importance of Predicate Sense

- added oracle predicate sense information as a feature in ASSERT
  - 60% of PropBanked Brown tagged with predicate sense information

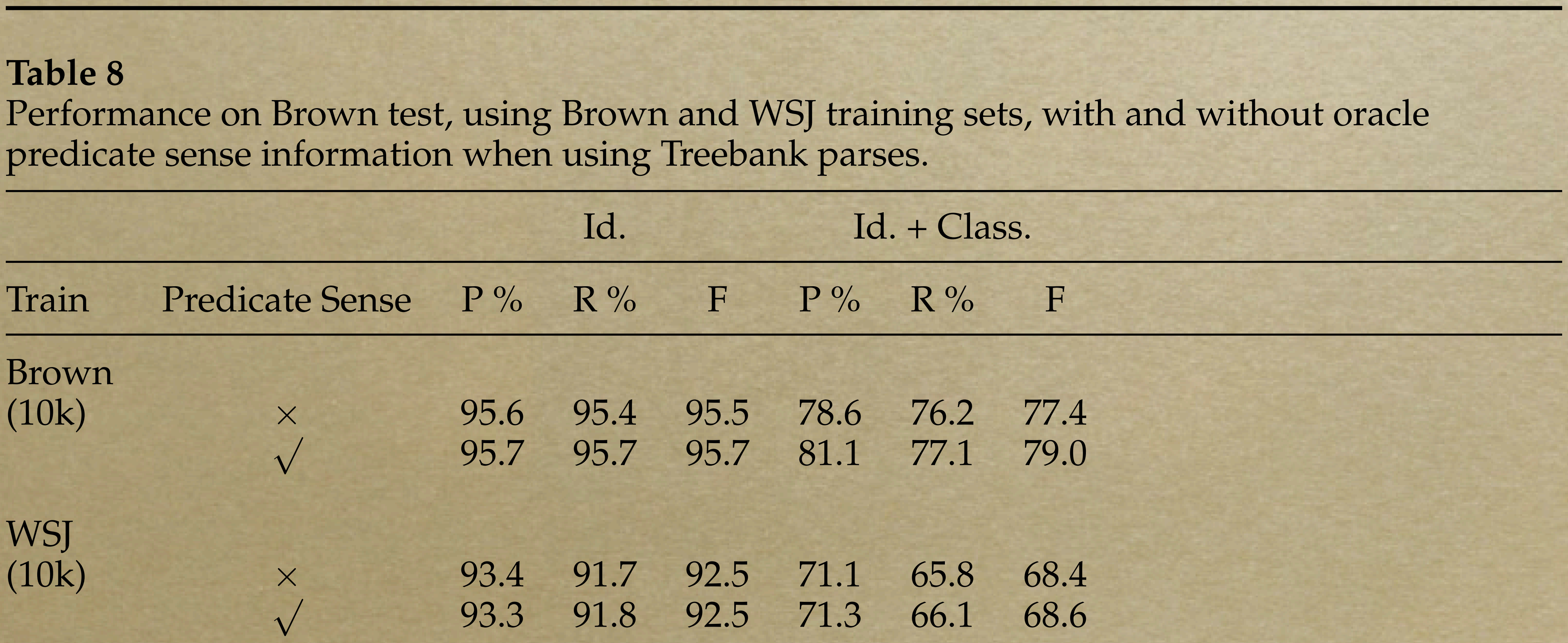


Table 8
Performance on Brown test, using Brown and WSJ training sets, with and without oracle predicate sense information when using Treebank parses.

		Id.			Id. + Class.		
Train	Predicate Sense	P %	R %	F	P %	R %	F
Brown (10k)	×	95.6	95.4	95.5	78.6	76.2	77.4
	√	95.7	95.7	95.7	81.1	77.1	79.0
WSJ (10k)	×	93.4	91.7	92.5	71.1	65.8	68.4
	√	93.3	91.8	92.5	71.3	66.1	68.6

Test Importance of Predicate Sense

- added oracle predicate sense information as a feature in ASSERT
  - 60% of PropBanked Brown tagged with predicate sense information

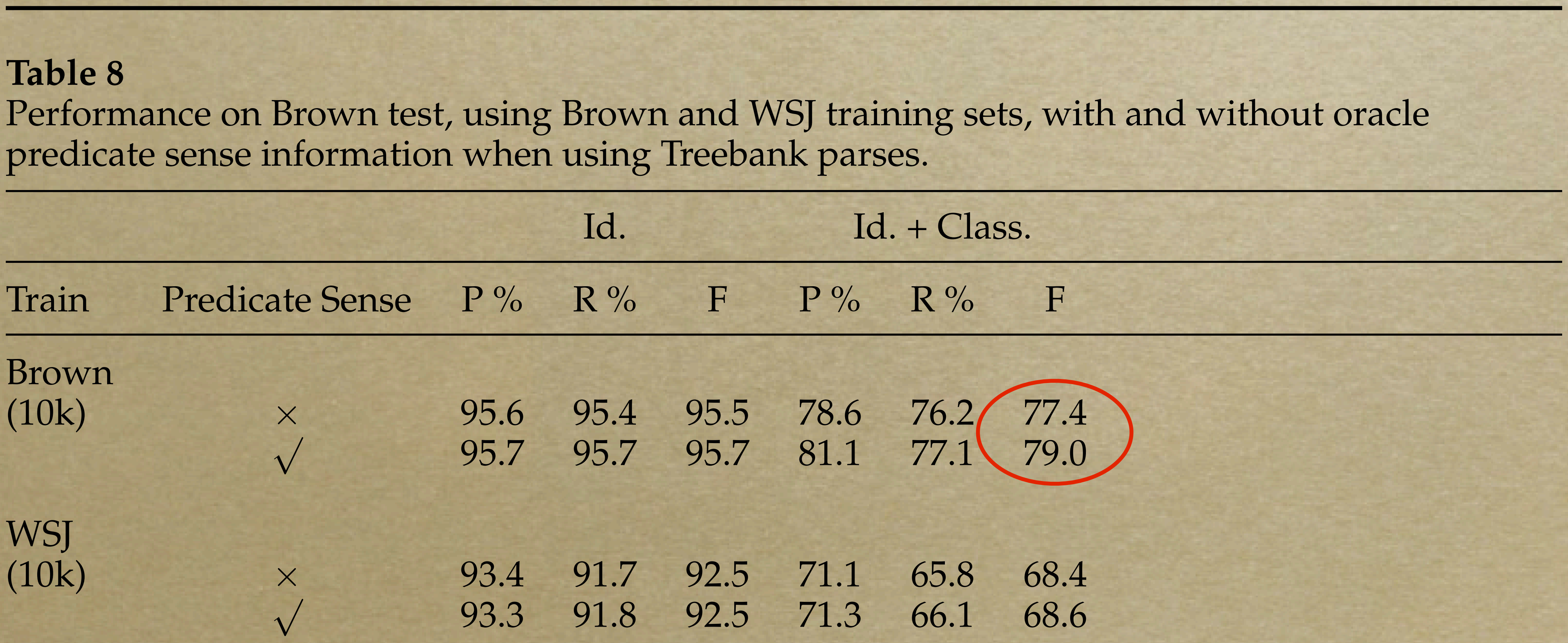


Table 8
Performance on Brown test, using Brown and WSJ training sets, with and without oracle predicate sense information when using Treebank parses.

		Id.			Id. + Class.		
Train	Predicate Sense	P %	R %	F	P %	R %	F
Brown (10k)	×	95.6	95.4	95.5	78.6	76.2	77.4
	√	95.7	95.7	95.7	81.1	77.1	79.0
WSJ (10k)	×	93.4	91.7	92.5	71.1	65.8	68.4
	√	93.3	91.8	92.5	71.3	66.1	68.6

Number of Unique Predicates and Head Words

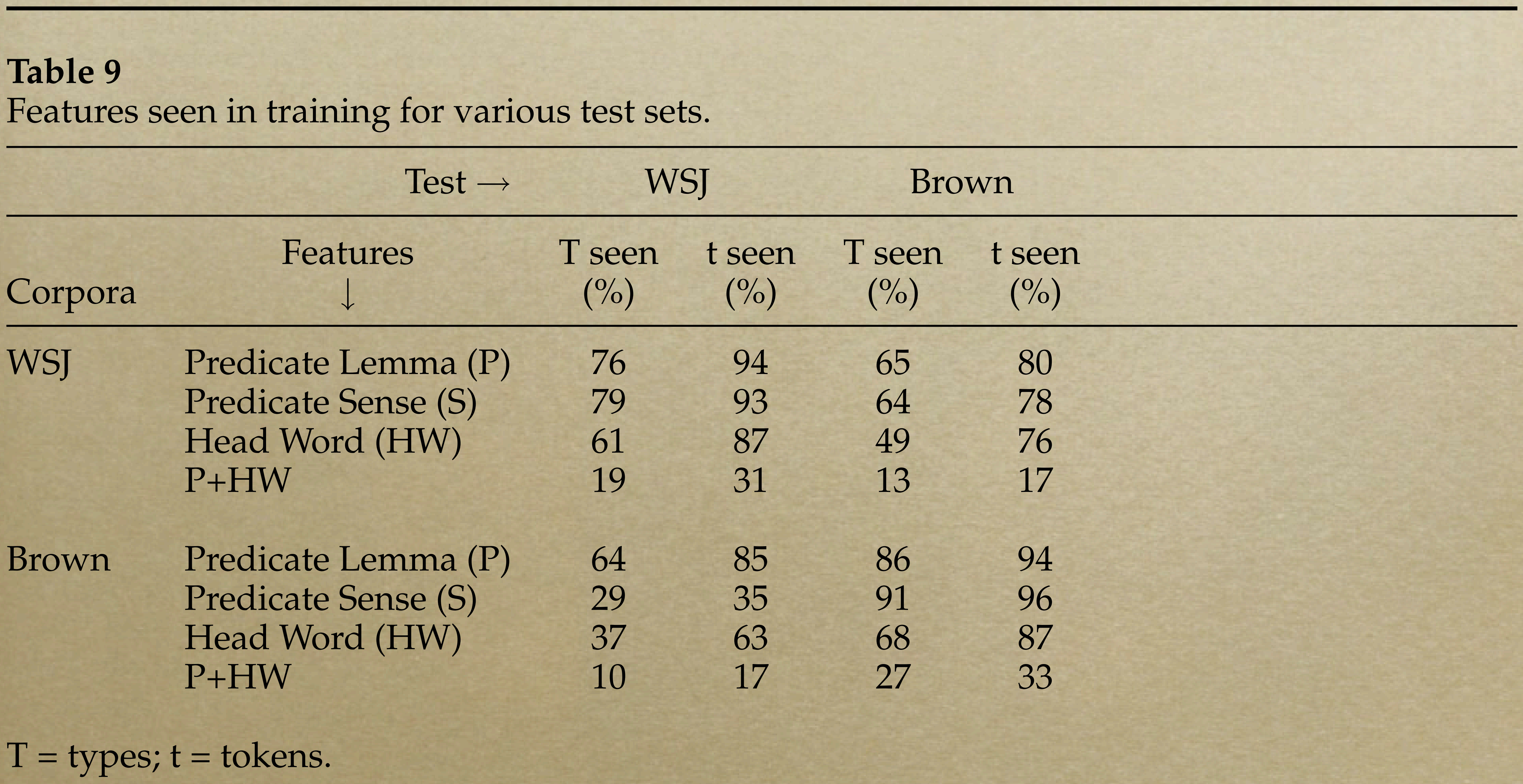


Table 9
Features seen in training for various test sets.

	Test →	WSJ		Brown	
Corpora	Features ↓	T seen (%)	t seen (%)	T seen (%)	t seen (%)
WSJ	Predicate Lemma (P)	76	94	65	80
	Predicate Sense (S)	79	93	64	78
	Head Word (HW)	61	87	49	76
	P+HW	19	31	13	17
Brown	Predicate Lemma (P)	64	85	86	94
	Predicate Sense (S)	29	35	91	96
	Head Word (HW)	37	63	68	87
	P+HW	10	17	27	33

T = types; t = tokens.

Number of Unique Predicates and Head Words

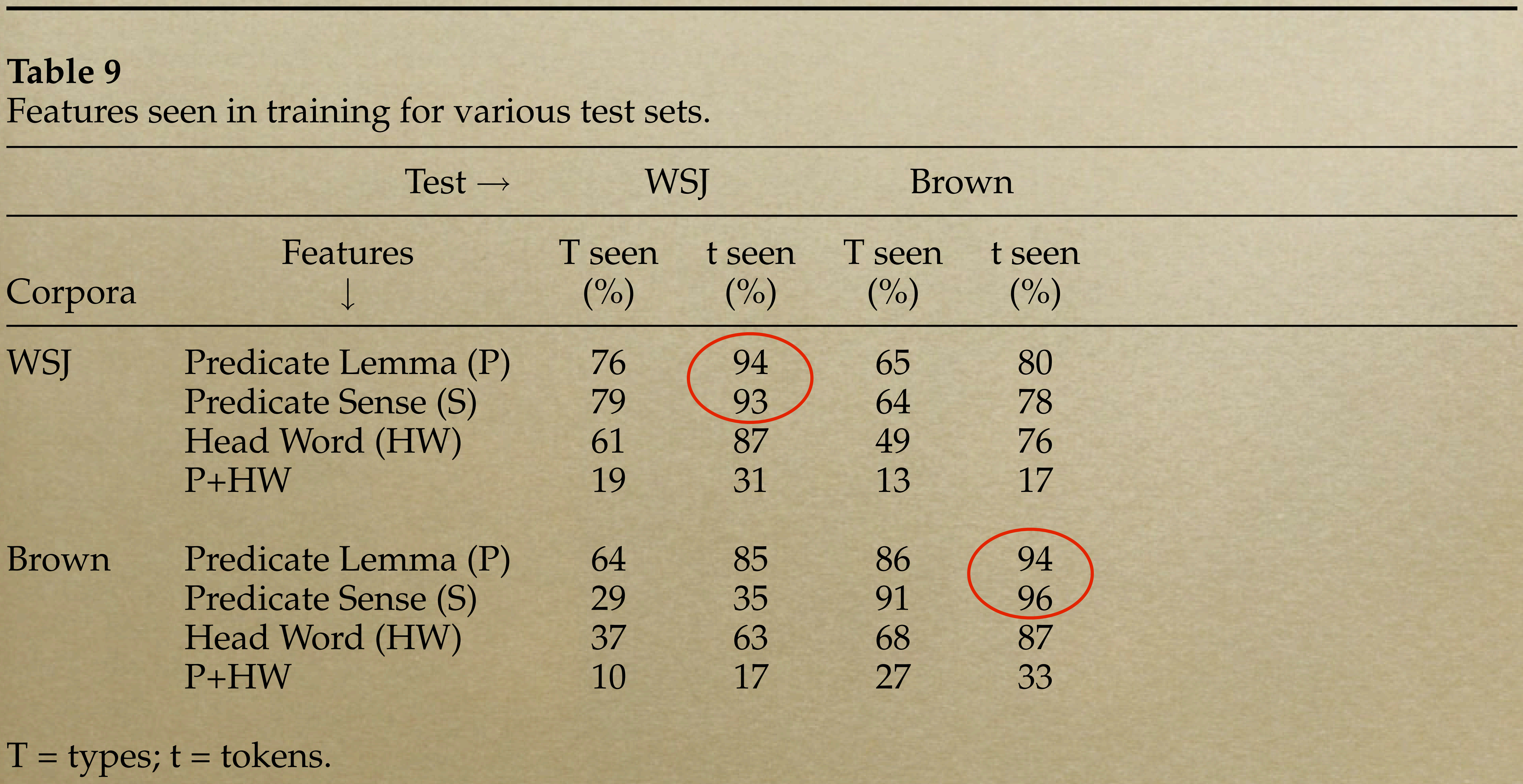


Table 9
Features seen in training for various test sets.

	Test →	WSJ		Brown	
Corpora	Features ↓	T seen (%)	t seen (%)	T seen (%)	t seen (%)
WSJ	Predicate Lemma (P)	76	94	65	80
	Predicate Sense (S)	79	93	64	78
	Head Word (HW)	61	87	49	76
	P+HW	19	31	13	17
Brown	Predicate Lemma (P)	64	85	86	94
	Predicate Sense (S)	29	35	91	96
	Head Word (HW)	37	63	68	87
	P+HW	10	17	27	33

T = types; t = tokens.

Number of Unique Predicates and Head Words

Table 9
Features seen in training for various test sets.

	Test →	WSJ		Brown	
Corpora	Features ↓	T seen (%)	t seen (%)	T seen (%)	t seen (%)
WSJ	Predicate Lemma (P)	76	94	65	80
	Predicate Sense (S)	79	93	64	78
	Head Word (HW)	61	87	49	76
	P+HW	19	31	13	17
Brown	Predicate Lemma (P)	64	85	86	94
	Predicate Sense (S)	29	35	91	96
	Head Word (HW)	37	63	68	87
	P+HW	10	17	27	33

T = types; t = tokens.

Number of Unique Predicates and Head Words

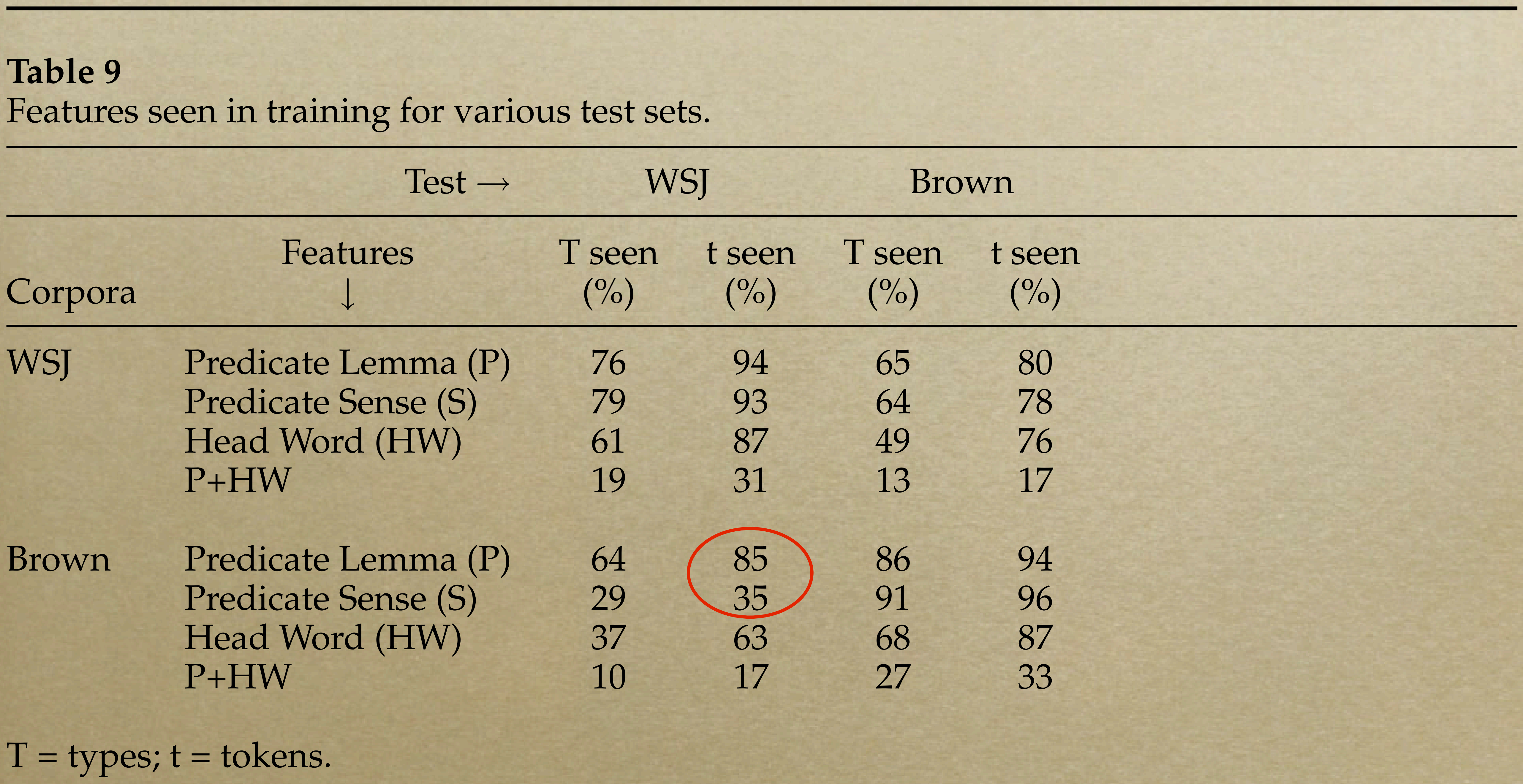


Table 9
Features seen in training for various test sets.

	Test →	WSJ		Brown	
Corpora	Features ↓	T seen (%)	t seen (%)	T seen (%)	t seen (%)
WSJ	Predicate Lemma (P)	76	94	65	80
	Predicate Sense (S)	79	93	64	78
	Head Word (HW)	61	87	49	76
	P+HW	19	31	13	17
Brown	Predicate Lemma (P)	64	85	86	94
	Predicate Sense (S)	29	35	91	96
	Head Word (HW)	37	63	68	87
	P+HW	10	17	27	33

T = types; t = tokens.

Distribution of PropBank Arguments

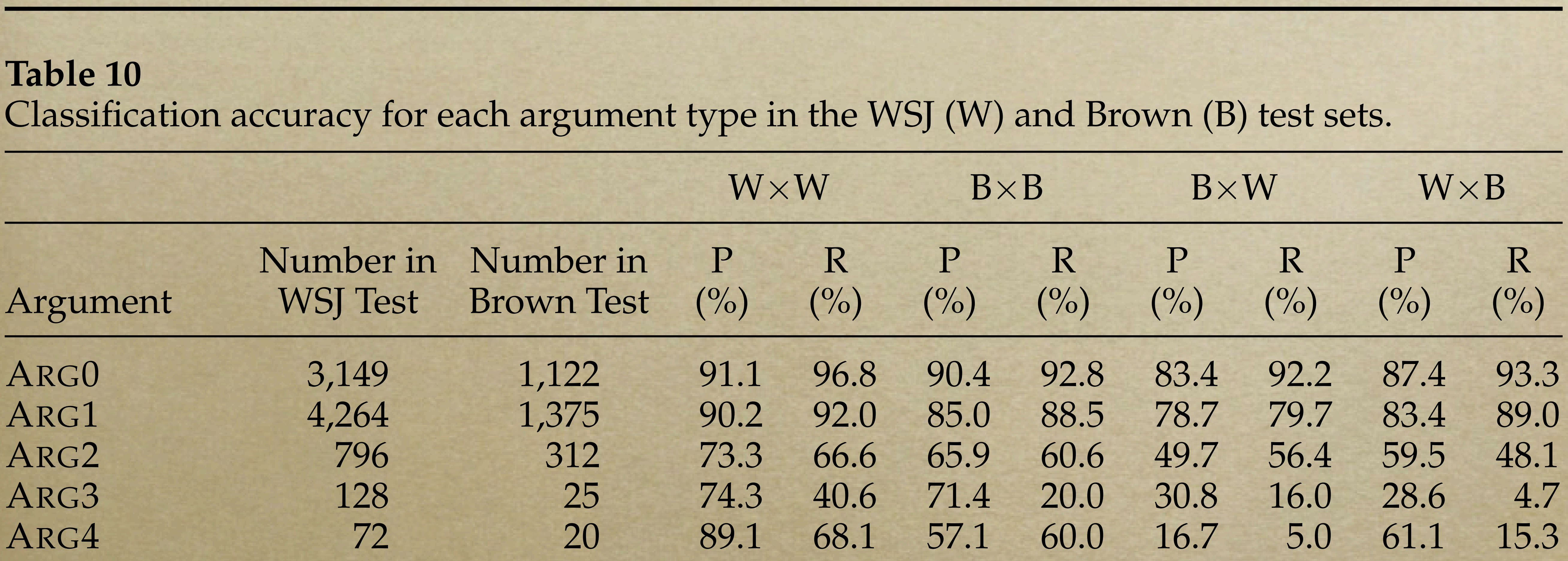


Table 10
Classification accuracy for each argument type in the WSJ (W) and Brown (B) test sets.

			W×W		B×B		B×W		W×B	
Argument	Number in WSJ Test	Number in Brown Test	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
ARG0	3,149	1,122	91.1	96.8	90.4	92.8	83.4	92.2	87.4	93.3
ARG1	4,264	1,375	90.2	92.0	85.0	88.5	78.7	79.7	83.4	89.0
ARG2	796	312	73.3	66.6	65.9	60.6	49.7	56.4	59.5	48.1
ARG3	128	25	74.3	40.6	71.4	20.0	30.8	16.0	28.6	4.7
ARG4	72	20	89.1	68.1	57.1	60.0	16.7	5.0	61.1	15.3

Distribution of PropBank Arguments

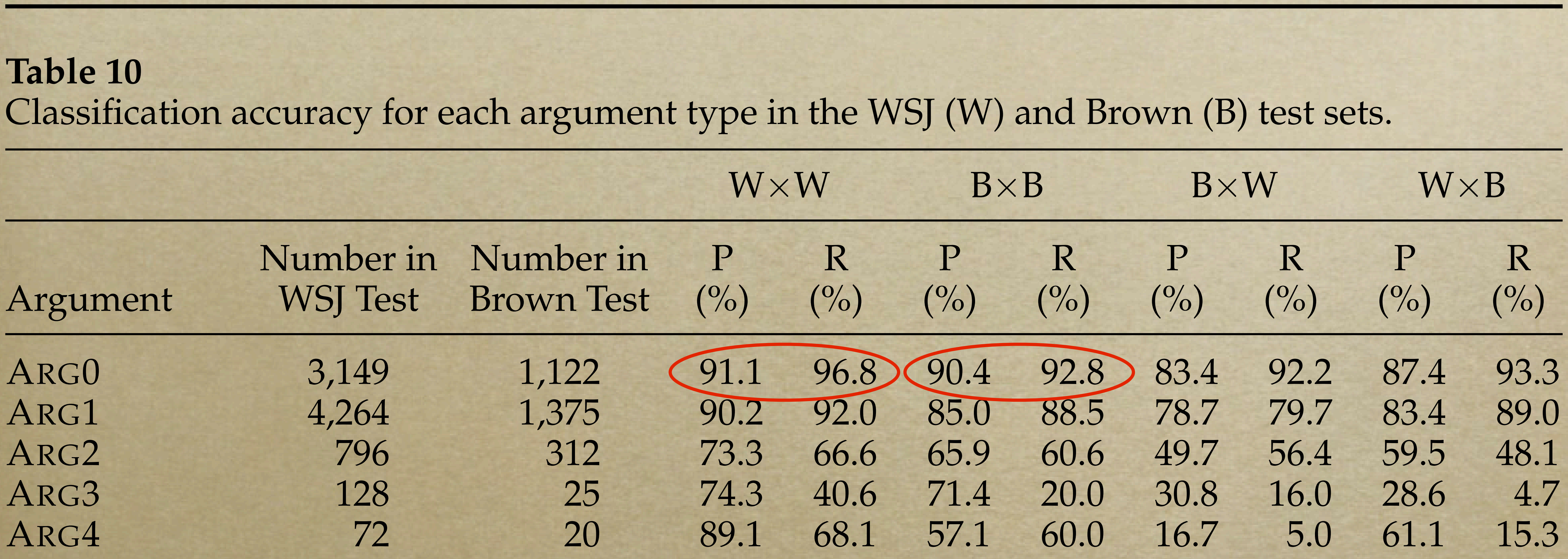


Table 10
Classification accuracy for each argument type in the WSJ (W) and Brown (B) test sets.

			W×W		B×B		B×W		W×B	
Argument	Number in WSJ Test	Number in Brown Test	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
ARG0	3,149	1,122	91.1	96.8	90.4	92.8	83.4	92.2	87.4	93.3
ARG1	4,264	1,375	90.2	92.0	85.0	88.5	78.7	79.7	83.4	89.0
ARG2	796	312	73.3	66.6	65.9	60.6	49.7	56.4	59.5	48.1
ARG3	128	25	74.3	40.6	71.4	20.0	30.8	16.0	28.6	4.7
ARG4	72	20	89.1	68.1	57.1	60.0	16.7	5.0	61.1	15.3

Distribution of PropBank Arguments

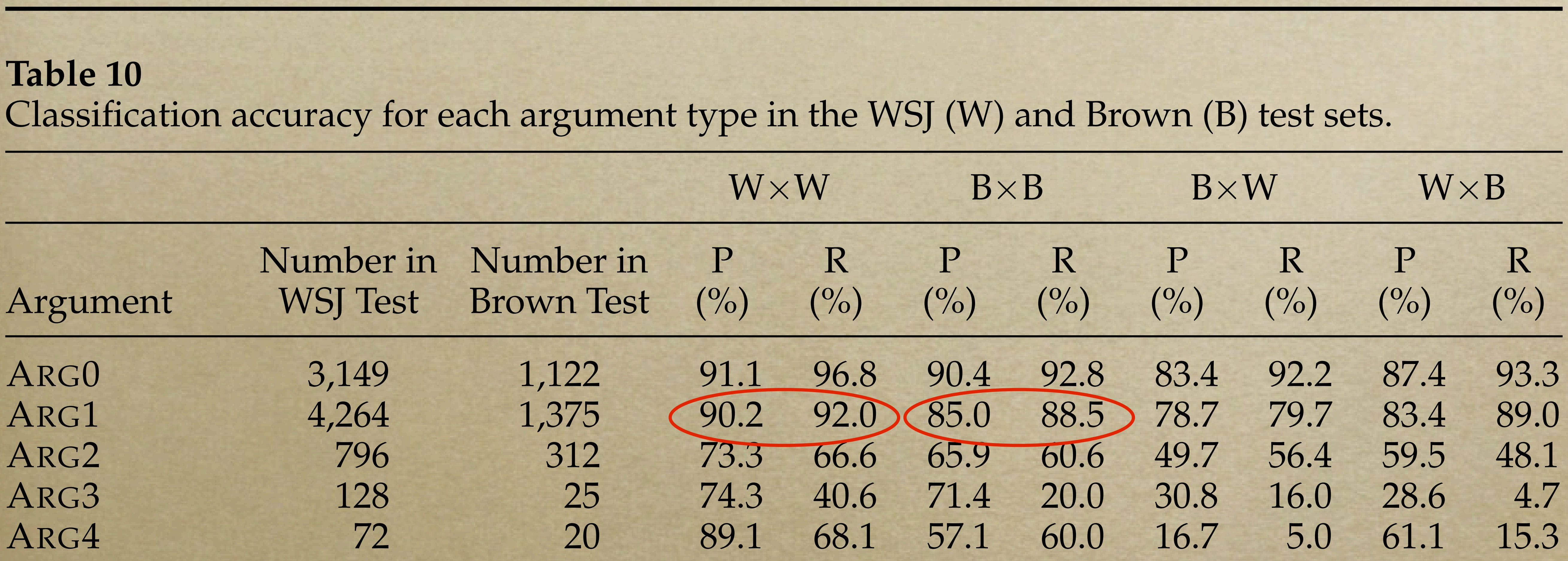


Table 10
Classification accuracy for each argument type in the WSJ (W) and Brown (B) test sets.

			W×W		B×B		B×W		W×B	
Argument	Number in WSJ Test	Number in Brown Test	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
ARG0	3,149	1,122	91.1	96.8	90.4	92.8	83.4	92.2	87.4	93.3
ARG1	4,264	1,375	90.2	92.0	85.0	88.5	78.7	79.7	83.4	89.0
ARG2	796	312	73.3	66.6	65.9	60.6	49.7	56.4	59.5	48.1
ARG3	128	25	74.3	40.6	71.4	20.0	30.8	16.0	28.6	4.7
ARG4	72	20	89.1	68.1	57.1	60.0	16.7	5.0	61.1	15.3

Distribution of PropBank Arguments

Table 10
Classification accuracy for each argument type in the WSJ (W) and Brown (B) test sets.

			W×W		B×B		B×W		W×B	
Argument	Number in WSJ Test	Number in Brown Test	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
ARG0	3,149	1,122	91.1	96.8	90.4	92.8	83.4	92.2	87.4	93.3
ARG1	4,264	1,375	90.2	92.0	85.0	88.5	78.7	79.7	83.4	89.0
ARG2	796	312	73.3	66.6	65.9	60.6	49.7	56.4	59.5	48.1
ARG3	128	25	74.3	40.6	71.4	20.0	30.8	16.0	28.6	4.7
ARG4	72	20	89.1	68.1	57.1	60.0	16.7	5.0	61.1	15.3

Distribution of Named Entities

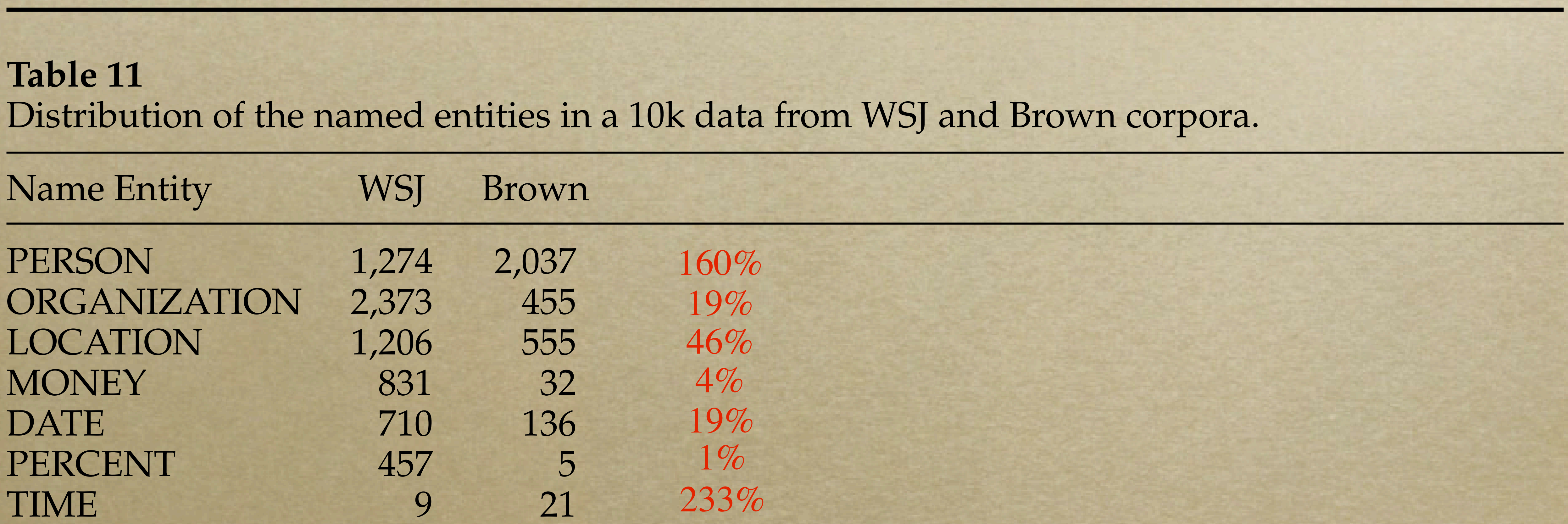


Table 11
Distribution of the named entities in a 10k data from WSJ and Brown corpora.

Name Entity	WSJ	Brown	
PERSON	1,274	2,037	160%
ORGANIZATION	2,373	455	19%
LOCATION	1,206	555	46%
MONEY	831	32	4%
DATE	710	136	19%
PERCENT	457	5	1%
TIME	9	21	233%

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

Effect of Improved Syntactic Parses

- recent improvements of Charniak parser available
    - n-best re-ranking
    - self-trained model using North American News corpus (NANC)
- five experiments conducted
    - used different versions of Charniak parser
    - train and test on WSJ and Brown

Effect of Improved Syntactic Parses (2)

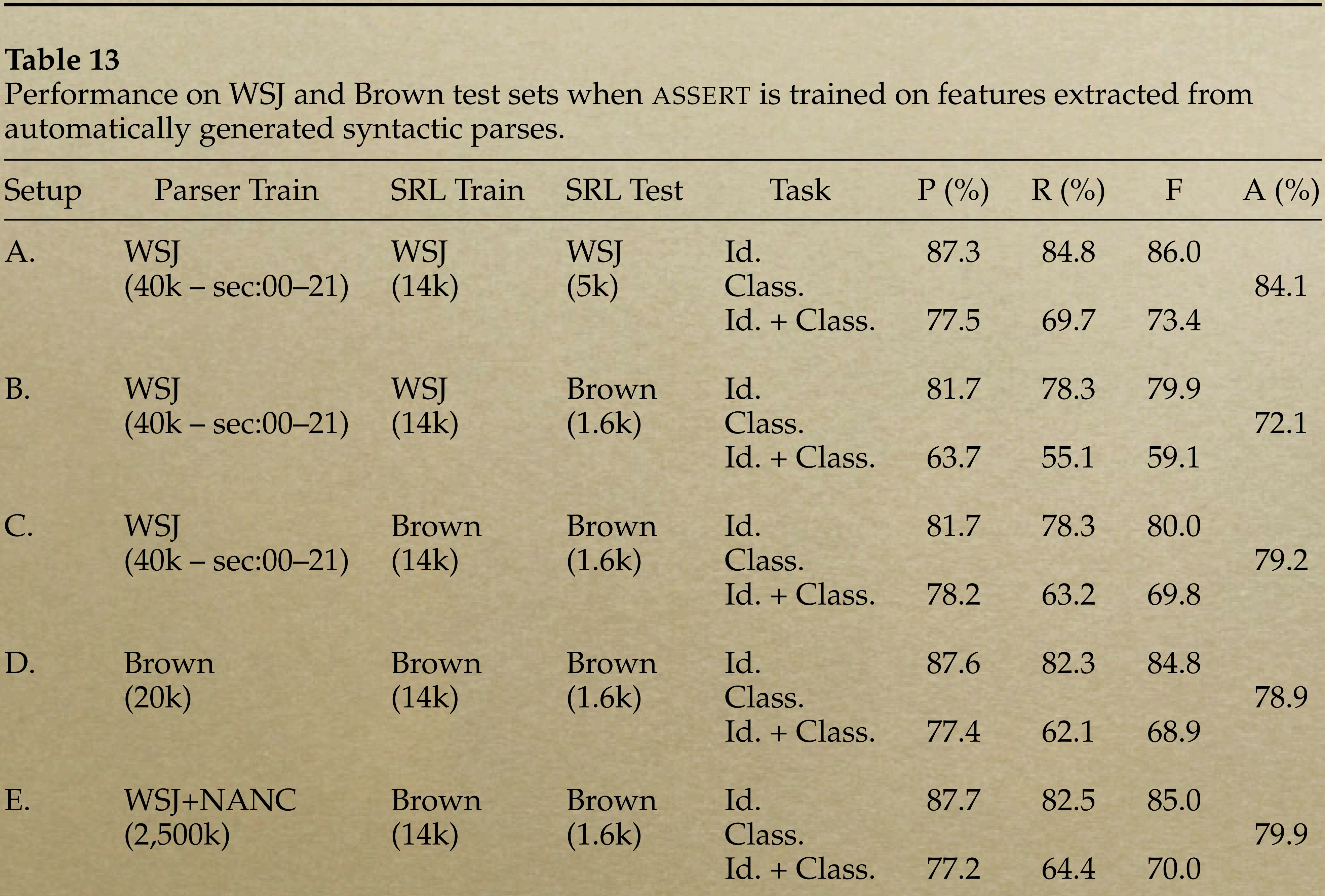


Table 13
Performance on WSJ and Brown test sets when ASSERT is trained on features extracted from automatically generated syntactic parses.

Performance for parser (f-score)	Setup	Parser Train	SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
91.0	A.	WSJ	WSJ	WSJ	Id.	87.3	84.8	86.0	
		(40k – sec:00–21)	(14k)	(5k)	Class.				84.1
					Id. + Class.	77.5	69.7	73.4	
91.0	B.	WSJ	WSJ	Brown	Id.	81.7	78.3	79.9	
		(40k – sec:00–21)	(14k)	(1.6k)	Class.				72.1
					Id. + Class.	63.7	55.1	59.1	
85.2	C.	WSJ	Brown	Brown	Id.	81.7	78.3	80.0	
		(40k – sec:00–21)	(14k)	(1.6k)	Class.				79.2
					Id. + Class.	78.2	63.2	69.8	
88.4	D.	Brown	Brown	Brown	Id.	87.6	82.3	84.8	
		(20k)	(14k)	(1.6k)	Class.				78.9
					Id. + Class.	77.4	62.1	68.9	
87.9	E.	WSJ+NANC	Brown	Brown	Id.	87.7	82.5	85.0	
		(2,500k)	(14k)	(1.6k)	Class.				79.9
					Id. + Class.	77.2	64.4	70.0	

Effect of Improved Syntactic Parses (2)

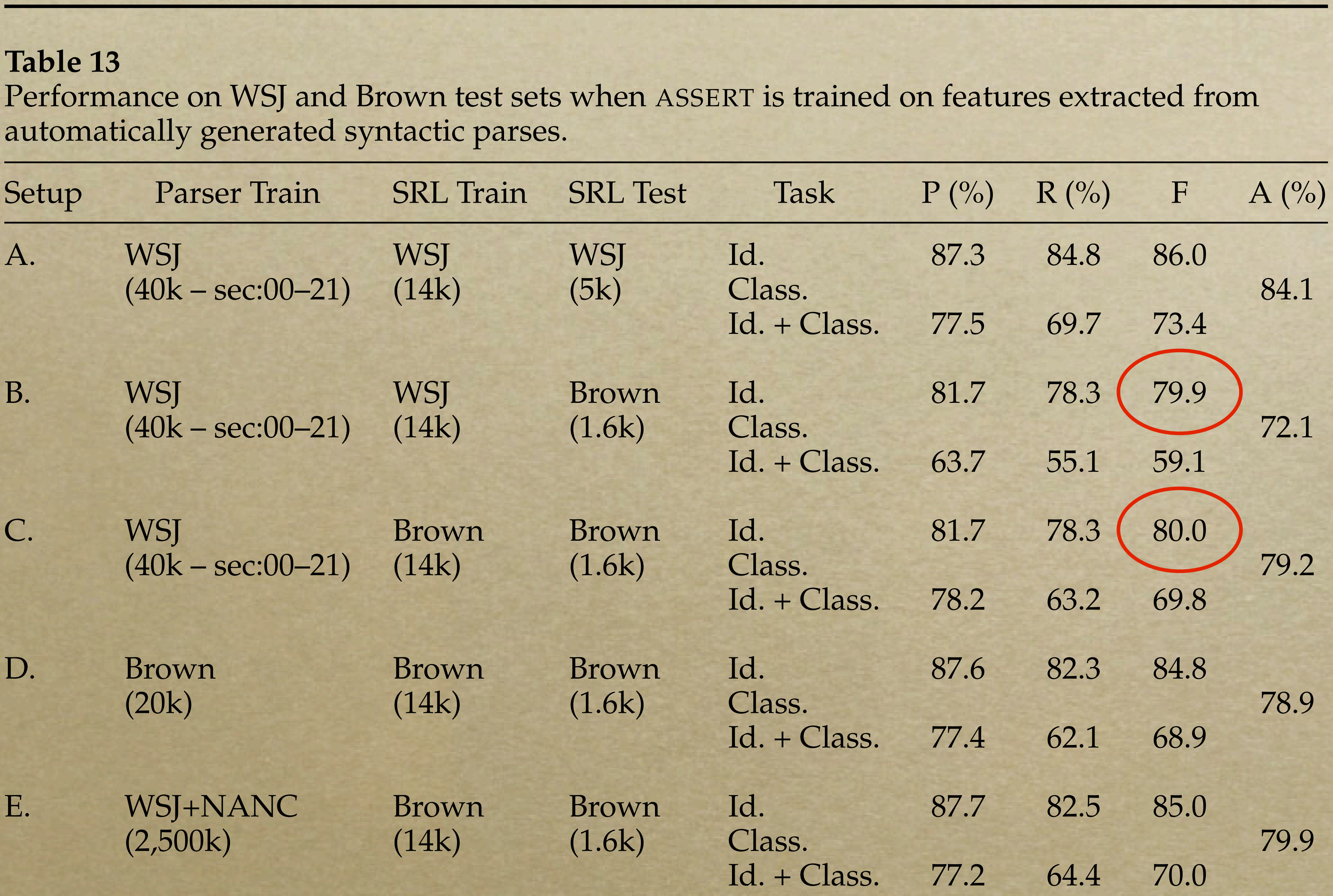


Table 13
Performance on WSJ and Brown test sets when ASSERT is trained on features extracted from automatically generated syntactic parses.

Performance for parser (f-score)	Setup	Parser Train	SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
91.0	A.	WSJ	WSJ	WSJ	Id.	87.3	84.8	86.0	
		(40k – sec:00–21)	(14k)	(5k)	Class.				84.1
					Id. + Class.	77.5	69.7	73.4	
91.0	B.	WSJ	WSJ	Brown	Id.	81.7	78.3	79.9	
		(40k – sec:00–21)	(14k)	(1.6k)	Class.				72.1
					Id. + Class.	63.7	55.1	59.1	
85.2	C.	WSJ	Brown	Brown	Id.	81.7	78.3	80.0	
		(40k – sec:00–21)	(14k)	(1.6k)	Class.				79.2
					Id. + Class.	78.2	63.2	69.8	
88.4	D.	Brown	Brown	Brown	Id.	87.6	82.3	84.8	
		(20k)	(14k)	(1.6k)	Class.				78.9
					Id. + Class.	77.4	62.1	68.9	
87.9	E.	WSJ+NANC	Brown	Brown	Id.	87.7	82.5	85.0	
		(2,500k)	(14k)	(1.6k)	Class.				79.9
					Id. + Class.	77.2	64.4	70.0	

Effect of Improved Syntactic Parses (2)

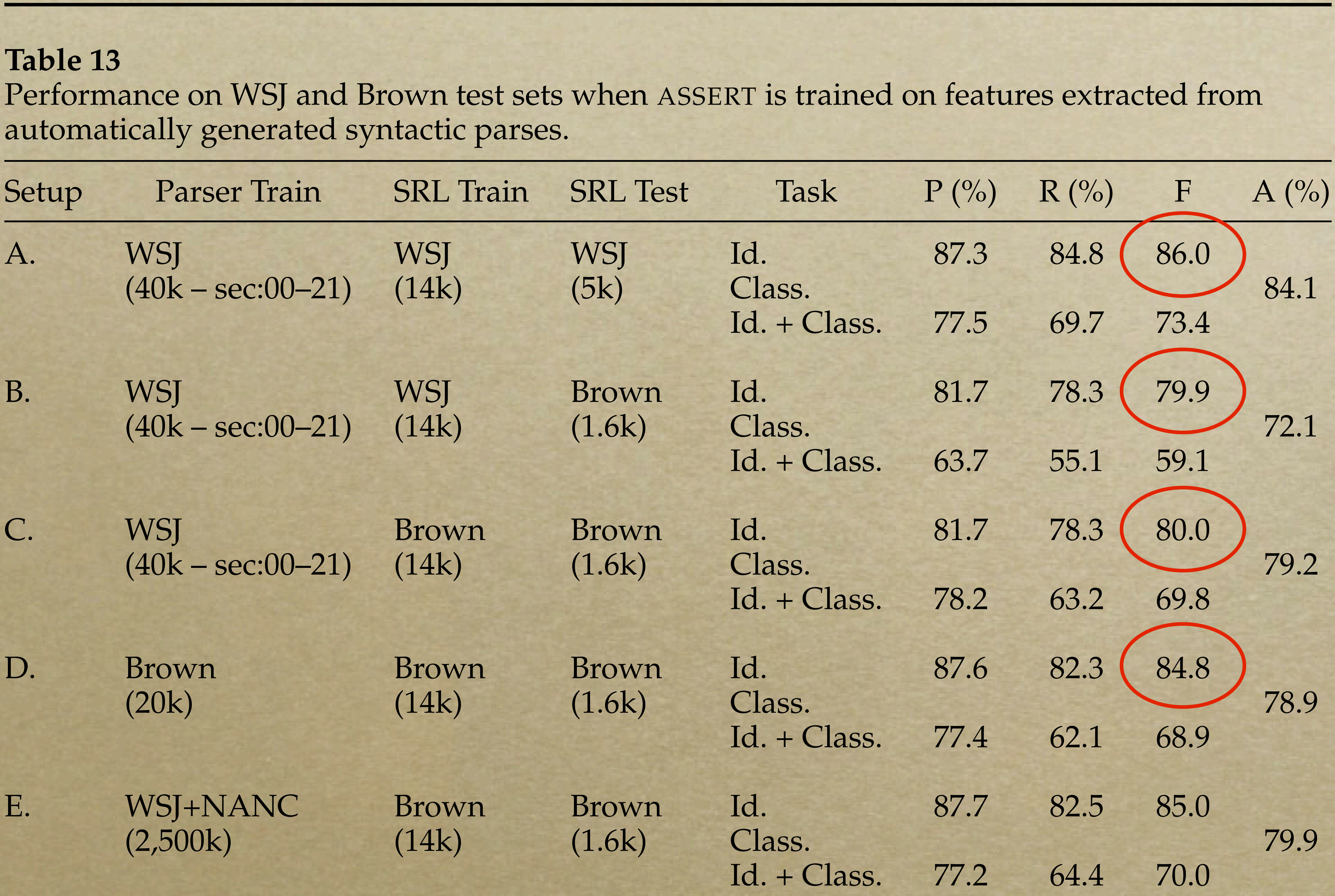


Table 13
Performance on WSJ and Brown test sets when ASSERT is trained on features extracted from automatically generated syntactic parses.

Performance for parser (f-score)	Setup	Parser Train	SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
91.0	A.	WSJ (40k – sec:00–21)	WSJ (14k)	WSJ (5k)	Id.	87.3	84.8	86.0	
					Class.				84.1
					Id. + Class.	77.5	69.7	73.4	
91.0	B.	WSJ (40k – sec:00–21)	WSJ (14k)	Brown (1.6k)	Id.	81.7	78.3	79.9	
					Class.				72.1
					Id. + Class.	63.7	55.1	59.1	
85.2	C.	WSJ (40k – sec:00–21)	Brown (14k)	Brown (1.6k)	Id.	81.7	78.3	80.0	
					Class.				79.2
					Id. + Class.	78.2	63.2	69.8	
88.4	D.	Brown (20k)	Brown (14k)	Brown (1.6k)	Id.	87.6	82.3	84.8	
					Class.				78.9
					Id. + Class.	77.4	62.1	68.9	
87.9	E.	WSJ+NANC (2,500k)	Brown (14k)	Brown (1.6k)	Id.	87.7	82.5	85.0	
					Class.				79.9
					Id. + Class.	77.2	64.4	70.0	

Effect of Improved Syntactic Parses (2)

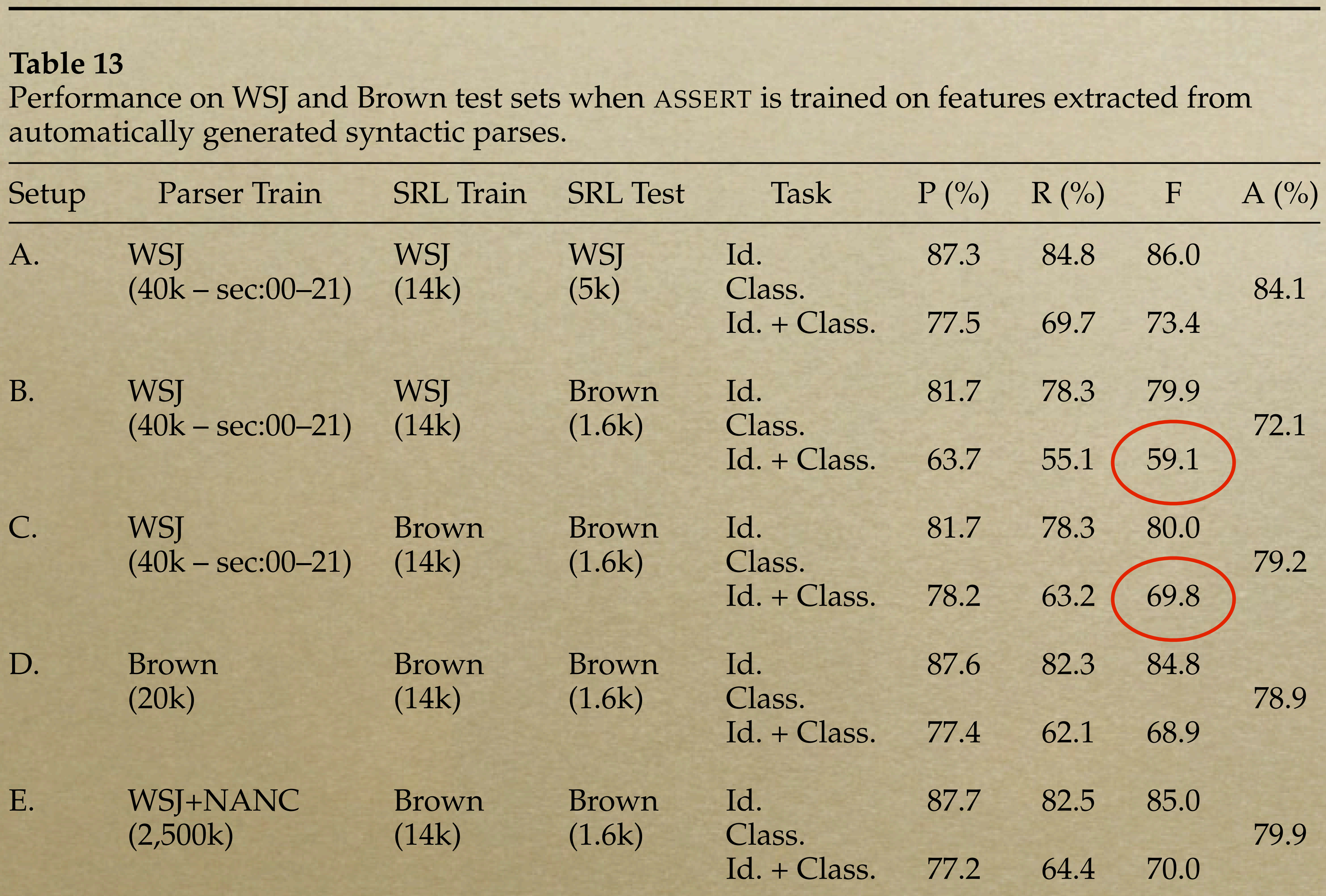


Table 13
Performance on WSJ and Brown test sets when ASSERT is trained on features extracted from automatically generated syntactic parses.

Performance for parser (f-score)	Setup	Parser Train	SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
91.0	A.	WSJ (40k – sec:00–21)	WSJ (14k)	WSJ (5k)	Id.	87.3	84.8	86.0	
					Class.				84.1
					Id. + Class.	77.5	69.7	73.4	
91.0	B.	WSJ (40k – sec:00–21)	WSJ (14k)	Brown (1.6k)	Id.	81.7	78.3	79.9	
					Class.				72.1
					Id. + Class.	63.7	55.1	59.1	
85.2	C.	WSJ (40k – sec:00–21)	Brown (14k)	Brown (1.6k)	Id.	81.7	78.3	80.0	
					Class.				79.2
					Id. + Class.	78.2	63.2	69.8	
88.4	D.	Brown (20k)	Brown (14k)	Brown (1.6k)	Id.	87.6	82.3	84.8	
					Class.				78.9
					Id. + Class.	77.4	62.1	68.9	
87.9	E.	WSJ+NANC (2,500k)	Brown (14k)	Brown (1.6k)	Id.	87.7	82.5	85.0	
					Class.				79.9
					Id. + Class.	77.2	64.4	70.0	

Effect of Improved Syntactic Parses (2)

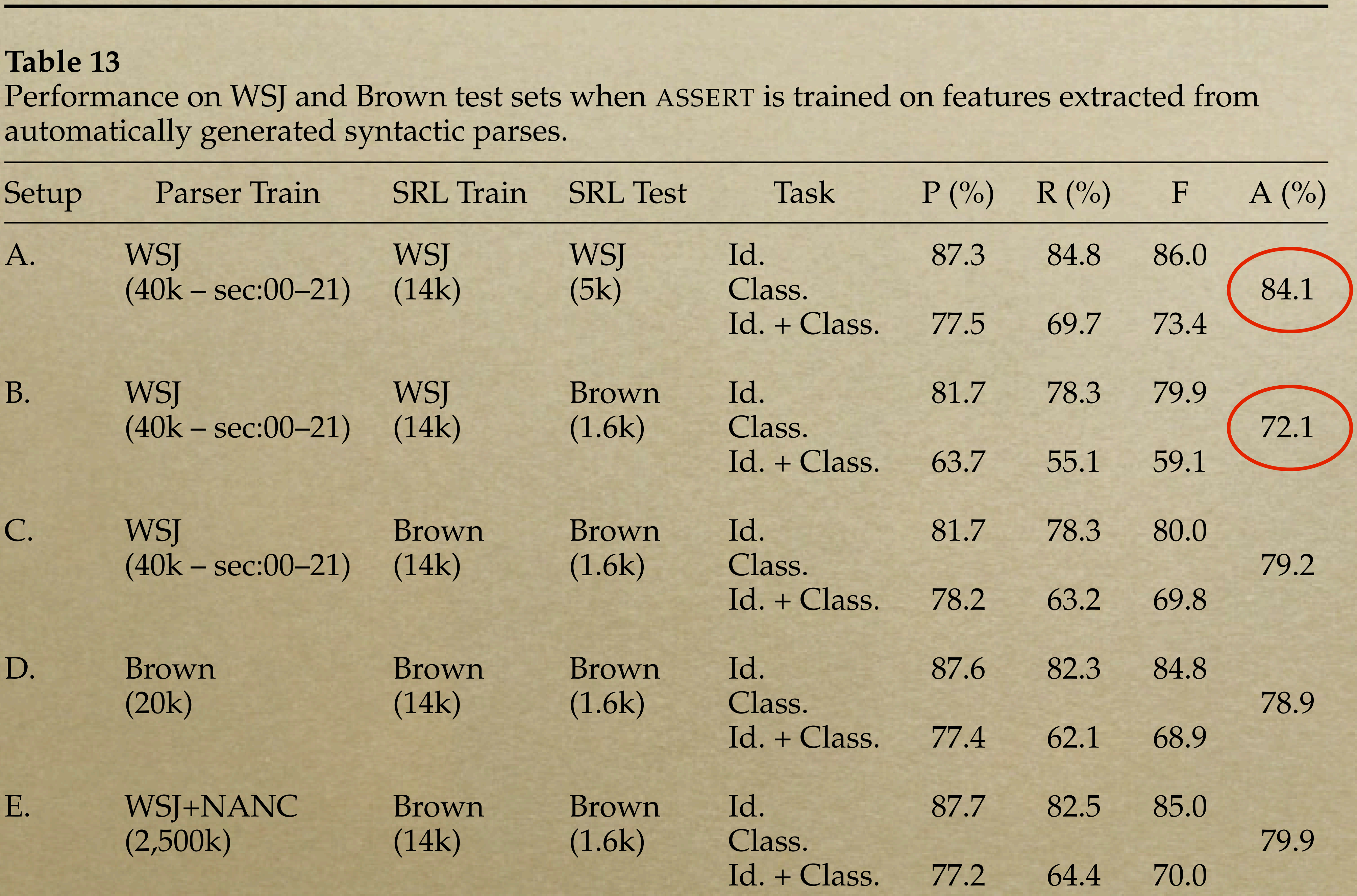


Table 13
Performance on WSJ and Brown test sets when ASSERT is trained on features extracted from automatically generated syntactic parses.

Performance for parser (f-score)	Setup	Parser Train	SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
91.0	A.	WSJ	WSJ	WSJ	Id.	87.3	84.8	86.0	
		(40k – sec:00–21)	(14k)	(5k)	Class.				84.1
					Id. + Class.	77.5	69.7	73.4	
91.0	B.	WSJ	WSJ	Brown	Id.	81.7	78.3	79.9	
		(40k – sec:00–21)	(14k)	(1.6k)	Class.				72.1
					Id. + Class.	63.7	55.1	59.1	
85.2	C.	WSJ	Brown	Brown	Id.	81.7	78.3	80.0	
		(40k – sec:00–21)	(14k)	(1.6k)	Class.				79.2
					Id. + Class.	78.2	63.2	69.8	
88.4	D.	Brown	Brown	Brown	Id.	87.6	82.3	84.8	
		(20k)	(14k)	(1.6k)	Class.				78.9
					Id. + Class.	77.4	62.1	68.9	
87.9	E.	WSJ+NANC	Brown	Brown	Id.	87.7	82.5	85.0	
		(2,500k)	(14k)	(1.6k)	Class.				79.9
					Id. + Class.	77.2	64.4	70.0	

Effect of Improved Syntactic Parses (2)

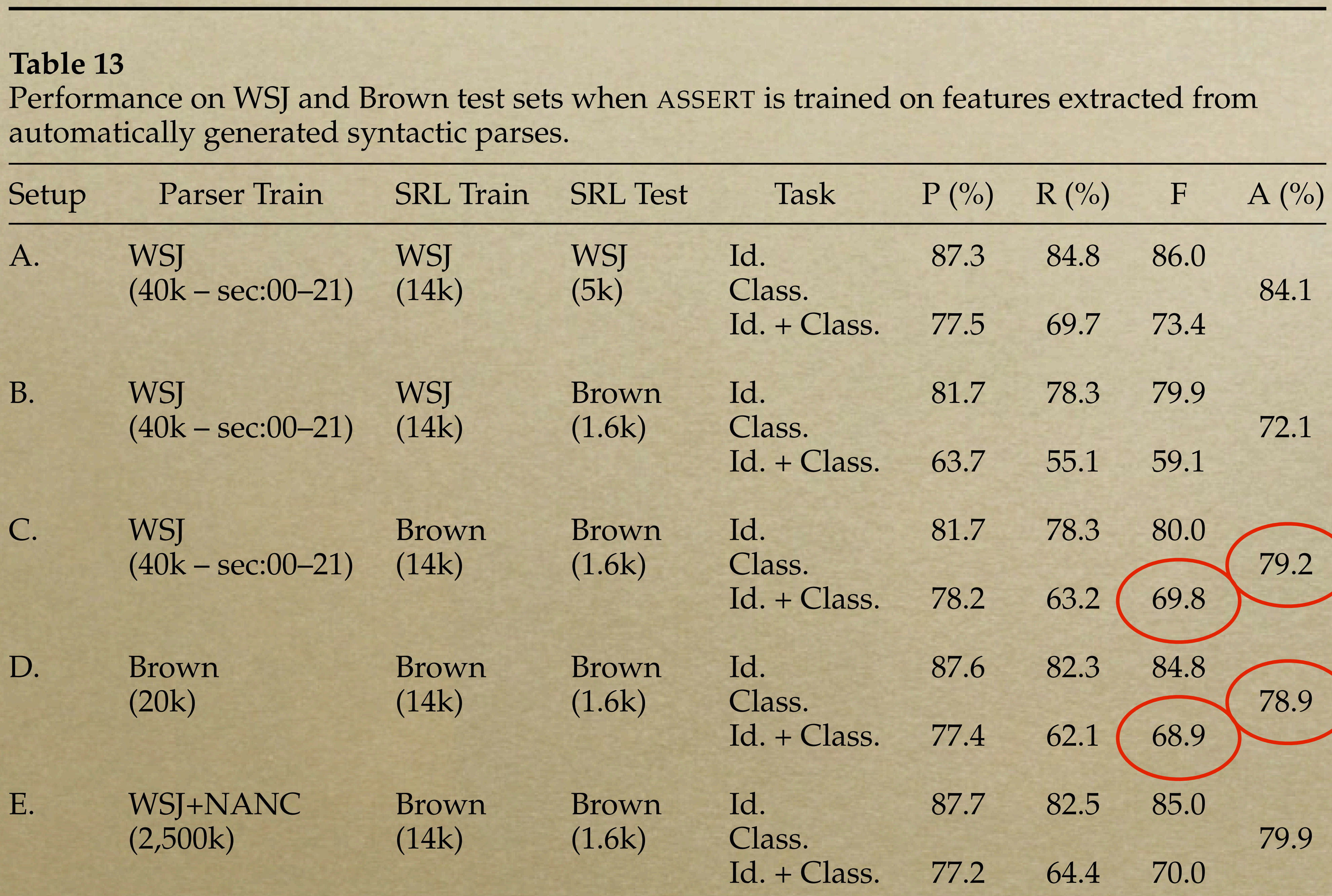


Table 13
Performance on WSJ and Brown test sets when ASSERT is trained on features extracted from automatically generated syntactic parses.

Performance for parser (f-score)	Setup	Parser Train	SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
91.0	A.	WSJ (40k – sec:00–21)	WSJ (14k)	WSJ (5k)	Id.	87.3	84.8	86.0	
					Class.				84.1
					Id. + Class.	77.5	69.7	73.4	
91.0	B.	WSJ (40k – sec:00–21)	WSJ (14k)	Brown (1.6k)	Id.	81.7	78.3	79.9	
					Class.				72.1
					Id. + Class.	63.7	55.1	59.1	
85.2	C.	WSJ (40k – sec:00–21)	Brown (14k)	Brown (1.6k)	Id.	81.7	78.3	80.0	
					Class.				79.2
					Id. + Class.	78.2	63.2	69.8	
88.4	D.	Brown (20k)	Brown (14k)	Brown (1.6k)	Id.	87.6	82.3	84.8	
					Class.				78.9
					Id. + Class.	77.4	62.1	68.9	
87.9	E.	WSJ+NANC (2,500k)	Brown (14k)	Brown (1.6k)	Id.	87.7	82.5	85.0	
					Class.				79.9
					Id. + Class.	77.2	64.4	70.0	

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

Adapting to a New Genre

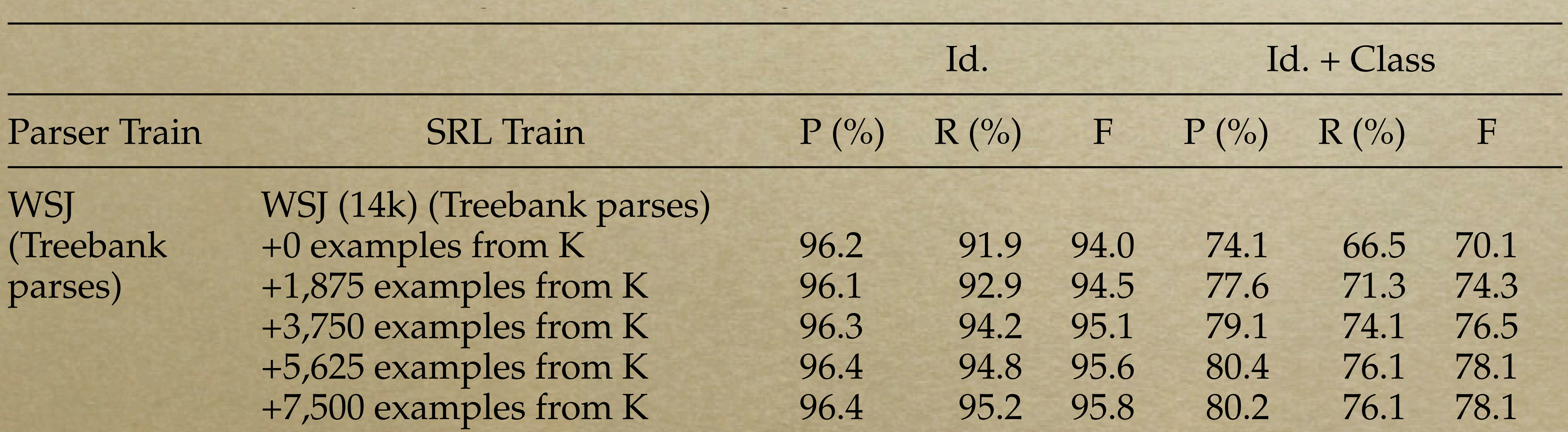
- incrementally add small amounts of data from new domain to out-of-domain training data
  - explore how much data is needed to achieve good results
- six scenarios
  - two use Treebank parses
  - four use different automatic parses
  - add predicates (0, 1875, 3750, 5625, 7500) from section K of Brown corpus

Adapting to a New Genre (2)

- for all six settings: performance for combined task improves gradually until about 5625 added examples from K, afterwards no/little improvement

Adapting to a New Genre (2)

- for all six settings: performance for combined task improves gradually until about 5625 added examples from K, afterwards no/little improvement



		Id.			Id. + Class		
Parser Train	SRL Train	P (%)	R (%)	F	P (%)	R (%)	F
WSJ (Treebank parses)	WSJ (14k) (Treebank parses)						
	+0 examples from K	96.2	91.9	94.0	74.1	66.5	70.1
	+1,875 examples from K	96.1	92.9	94.5	77.6	71.3	74.3
	+3,750 examples from K	96.3	94.2	95.1	79.1	74.1	76.5
	+5,625 examples from K	96.4	94.8	95.6	80.4	76.1	78.1
	+7,500 examples from K	96.4	95.2	95.8	80.2	76.1	78.1

Adapting to a New Genre (2)

- for all six settings: performance for combined task improves gradually until about 5625 added examples from K, afterwards no/little improvement
- even when parser is trained on WSJ and SRL is trained on WSJ, adding 7500 instances of the new genre achieves almost the same performance as when all three are from Brown (67.2 vs. 69.9)
- for identification task: little improvement when adding examples from new genre

Overview

- Introduction
- Corpora and Task Description
- ASSERT
- Robustness Experiments
- Effect of Improved Syntactic Parses
- Adapting to a New Genre
- Conclusion

Conclusions

- for SRL trained on WSJ data, the system‘s performance on the Brown test set drops largely compared to WSJ test data
- major performance loss occurs in classification task, identification task is only responsible for relatively small drop
- errors in syntactic parser are not a large factor in the overall performance difference

Conclusions (2)

- final hypothesis: **The Brown corpus is in some sense fundamentally more difficult for SRL problems.**
  - Brown is a more heterogeneous source than WSJ
- more homogenous training data allows the system to rely heavily on specific features and relations
- usually more heterogeneous data ports better to other corpora

Conclusions (3)

- two main possibilities to improve performance for cross-genre classification
  - use less homogenous corpora => draw fewer examples from many sources rather than using many examples from one source
  - use less specific features => reduce likelihood of learning idiosyncratic aspects of training domain
- probably reduce performance for same-genre classification, but improve performance for cross-genre classification

Thank you for your attention!

Questions?