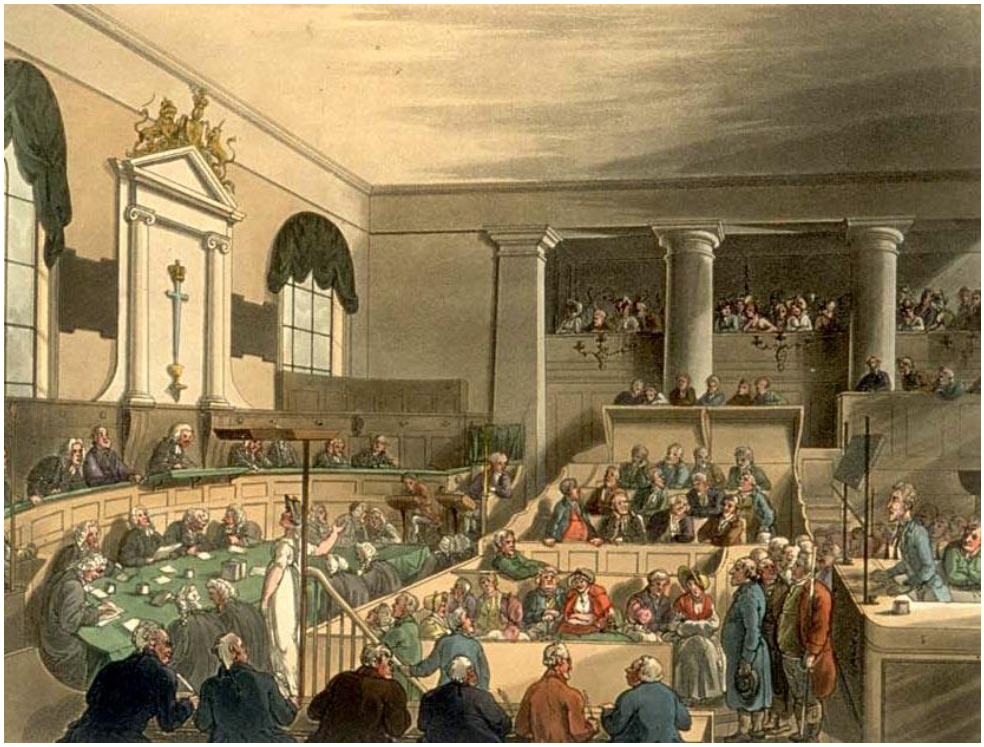

The Old Bailey Corpus 2.0, 1720-1913

Manual

Magnus Huber, Magnus Nissel & Karin Puga

2016-05-01

DFG Deutsche Forschungsgemeinschaft

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

CLARIN-D

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

UNIVERSITÄT DES SAARLANDES

CLARIN CENTRE B

Contact

Magnus Huber
Department of English, FB 05
Justus Liebig University Giessen
Otto-Behaghel-Str. 10 B
D-35394 Giessen
Germany
magnus.huber@anglistik.uni-giessen.de

Contents

Cover illustration: coloured aquatint by Thomas Rowlandson and Auguste Charles Pugin of a trial at the Old Bailey published in

Ackermann, Rudolph & Pyne, William Henry. 1808-1810. *The microcosm of London or London in miniature.* Vol. II. London: Rudolph Ackermann, plate facing p. 212.

William Pyne describes the scene as follows: "The plate represents the court employed in the examination of a witness, who appears to have just received the usual admonition upon these occasions, of '*Hold up your head, young woman, and look at his lordship*'" (Ackermann & Pyne 108-1810: 212)

1. Acknowledgements

The *Old Bailey Corpus* (*OBC*) is the fruit of over ten years of work, which could not have been achieved out without the generous support of the following institutions and individuals:

Thanks are due to Justus Liebig University Giessen for a start-up grant and a project room that enabled us to compile the *OBC* from 2005 on. We also gratefully acknowledge the major financial support of the German Science Foundation from May 2008 to January 2012 (DFG, HU 884/6-1, HU 884/6-2). Version 2.0 of the corpus has now been integrated into the German section of the Common Language Resources and Technology Infrastructure (CLARIN-D) to achieve persistent storage and access. This integration was generously funded by the German Federal Ministry of Education and Research from January 2015 to March 2016 (CLARIN-DE-FAG2-KP3) and coordinated by CLARIN-D.

The *OBC* is based on the *Proceedings of the Old Bailey*. We are indebted to the creators of this resource for kindly providing us with digitalized transcripts of the *Proceedings* and for discussing technical and content-related aspects with us: Tim Hitchcock (Department of History, University of Sussex), Robert Shoemaker (Department of History, Humanities Research Institute, University of Sheffield) and Sharon Howard (Department of History, Humanities Research Institute, University of Sheffield) were always generous with their data and advice.

For welcoming the *OBC* to the CLARIN-D framework and for his constant support and encouragement during the integration phase we would like to thank the head of the CLARIN-D Working Group 2 "Other Philologies", Christian Mair. Also, the adminsitrative support of Christian Drude, Julia Misersky of the Max-Planck Institute for Psycholinguistics in Nijmegen and of Gregor Wiedemann of the University of Leipzig is greatly appreciated. Last but not least, our heartfelt thanks go to Elke Teich, Hannah Kermes and Jörg Knappen of Saarland University and Peter Fankhauser of the Institute of the German Language in Mannheim for their untiring technical support and for agreeing to host the *OBC* at the CLARIN-D Service Centre Saarbrücken.

Giessen, 1 May 2016

Magnus Huber, Magnus Nissel and Karin Puga

2. Overview

The *Old Bailey Corpus* (*OBC*) is a sociolinguistically, pragmatically and textually annotated corpus based on a selection of the *Proceedings of the Old Bailey* (henceforth *Proceedings*; Hitchcock et al. 2015), the published version of the trials at London's Central Criminal Court. For an electronic version of the *Proceedings* as well as detailed background information on the history of their publication as well as that of the Old Bailey, consult the excellent *Old Bailey Proceedings Online*.

The 2,163 volumes of the *Proceedings* contain almost 200,000 trials, totalling ca. 134 million words. These speech-related texts record Late Modern English as spoken in the courtroom. The Old Bailey trial proceedings were taken down in shorthand and as such the published *Proceedings* are a reasonably close approximation of what was said in court, even though scribes, printers, publishers and the constraints of the printed medium acted as linguistic filters between the spoken word and its representation in the *Proceedings*.

The compilation of the *OBC* started in January 2006 at Justus Liebig University Giessen. Version 1.0 of the *OBC* was released in 2013, containing 14 million words. Turning the digitalized *Proceedings* into the linguistic *OBC* consisted of four main steps:

1. Selection of a balanced subset of the *Proceedings* to achieve a roughly equal amount of spoken words per decade,
2. Automatic identification and tagging of utterances in the *Proceedings* with the help of Perl and Python scripts,
3. Sociolinguistic, pragmatic and textual annotation of every utterance, based on sociobiographical speaker data found in the context of the trials,
4. Part-of-speech tagging of the *Proceedings* using the CLAWS 7 tagset.

From January 2015 to March 2016, *OBC 2.0* (10 million words larger than version 1.0) was integrated into the German section of the Common Language Resources and Technology Infrastructure (CLARIN-D) to achieve persistent storage and access. *OBC 2.0* can now be accessed online via CQPweb at the Saarland University CLARIN-D repository. The corpus can also be downloaded from the Old Bailey Landing Page together with a custom-made search tool.

Version 2.0 of the *OBC* consists of 637 selected *Proceedings*, from 1720 to 1913. *OBC 2.0* contains a total of 24.4 million words, with 1.2 million speech-related words per decade on average. Three periods have a noticeably lower number of spoken words, 1720-1729 (71,185 words), 1730-1739 (938,902 words) and 1910-1913 (710,914 words). All *Proceedings* from the 1720s and 1730s were included in *OBC 2.0*, but there are only relatively few verbatim reports in the *Proceedings* of the 1720s and only just under 1 million spoken words in the 1730s. The publication of the *Proceedings* was discontinued in 1913, so this last "decade" contains just four years and accordingly only 710,914 words were included here.

OBC 2.0 allows the linguist to analyze speech-related texts in a period that has been neglected both with regard to the compilation of primary linguistic data and the description of the structure, variability, and change of spoken English. With a high number of speakers and over half a million individual utterances, *OBC 2.0* constitutes a fairly representative sample of spoken, rather formal Late Modern English in the courtroom setting. Moreover, every speaker turn is annotated for sociobiographical (gender, social class, age), pragmatic (role in the trial) and textual variables (the shorthand scribe, printer and publisher of individual *Proceedings*).

OBC 2.0 is the largest diachronic collection of spoken English with this detail of utterance level sociolinguistic annotation. Although the corpus can of course be used for traditional investigations of language change, it is particularly suited for studies that correlate linguistic change and structural variability in Late Modern English with the social context. Its size, the time span covered (almost 200 years) and the available sociobiographical speaker information make *OBC 2.0* ideal for fine-tuned studies involving several independent variables, including historical sociolinguistic approachesand the analysis of low-frequency features.

3. How to access the *Old Bailey Corpus 2.0*

There are two ways to access the Old Bailey Corpus:

- Via CQPweb at the CLARIN-D Service Centre of Saarland University (requires free registration).
- Via a downloadable version, together with a custom-made concordancer, available from the Old Bailey Landing Page.

4. Structure of the corpus: file format, file names, text structure and word counts

File format

OBC 2.0 consists of 637 files (individual *Proceedings*) in Extensible Markup Language (XML) that are tagged according to the guidelines of the Text Encoding Initiative (TEI).

File names

The file names of the individual *Proceedings* are composed of the name and version of the corpus, followed by the year, month and day of its original publication in the format yyyymmdd. For example, OBC2-17200427 is the *Proceeding* published on 27 April 1720 as included in *OBC 2.0* and OBC2POS-17200427 is the part-of-speech tagged version of the same *Proceeding*.

Text structure

OBC 2.0 consists of 637 different files (*Proceedings*), amounting to 517,769 utterances and 24,443,588 spoken words. The files for *OBC 2.0* were selected using stratified random sampling to arrive at a relatively balanced number of spoken words per decade. The *Proceedings* included in the corpus are listed in the word count spreadsheet available for download on the Old Bailey Landing Page.

The title page of each *Proceeding* states the dates of the sessions and begins with a list of the judges and other court personnel, as well as the names of the jurors. This is followed by a sequence of trials of varying number and length. The trials are usually introduced by an indictment, which is followed by the trial proper, consisting of witness statements and cross-examinations and, finally, a verdict. This structure is reflected in the XML markup of the digitalized transcripts of the *Proceedings* obtained from Tim Hitchcock (Department of History, University of Sussex), Robert Shoemaker (Department of History, Humanities Research Institute, University of Sheffield) and Sharon Howard (Department of History, Humanities Research Institute, University of Sheffield). To make *OBC 2.0* as versatile as possible, the original markup of the digitalized *Proceedings* was retained, with only a few minor adaptions (see Section 5).

Word counts

For the purpose of word counts in *OBC 2.0*, a word is defined as an uninterrupted string of characters, excluding apostrophes and hyphens, and delimited by punctuation or white space. The word counts are based on the CLAWS POS-tagged version of *OBC 2.0*, meaning that synthetic genitives such as *doctor's* or contracted forms such as *can't* are counted as two words

(since CLAWS transforms these into *doctor_NN1 's_GE* and *ca_VM n't_XX*). However, obsolete spellings of past and past participle forms involving apostropes such as *cry't* for *cried*, *depos'd* for *deposed* or *before-mention'd* for *before-mentioned*, counted as one word.

For purposes of normalization, word counts for individual *Proceedings* as well as decades are available in a spreadsheet that can be downloaded from the Old Bailey Landing Page. Word counts are available for the following categories:

Total number of words

- N words total, N utterances, N words spoken

Gender

- female, male, unknown

Class

- higher (HISCO 1-5), lower (HISCO 6-13), unknown

Class x Gender

- higher: females, males, unknown
- lower: females, males, unknown

Age

- known, unknown

Speaker role

- judge, lawyer, victim, defendant, witness, interpreter, unknown

Scribe, Editor, Printer, Publisher

- of individual *Proceedings*

Diagram 1 gives an overview of the number of spoken words in 40-year periods in *OBC 2.0* und indicates for how many words in each period the role in the courtroom, the gender and the social class of the speaker is known:

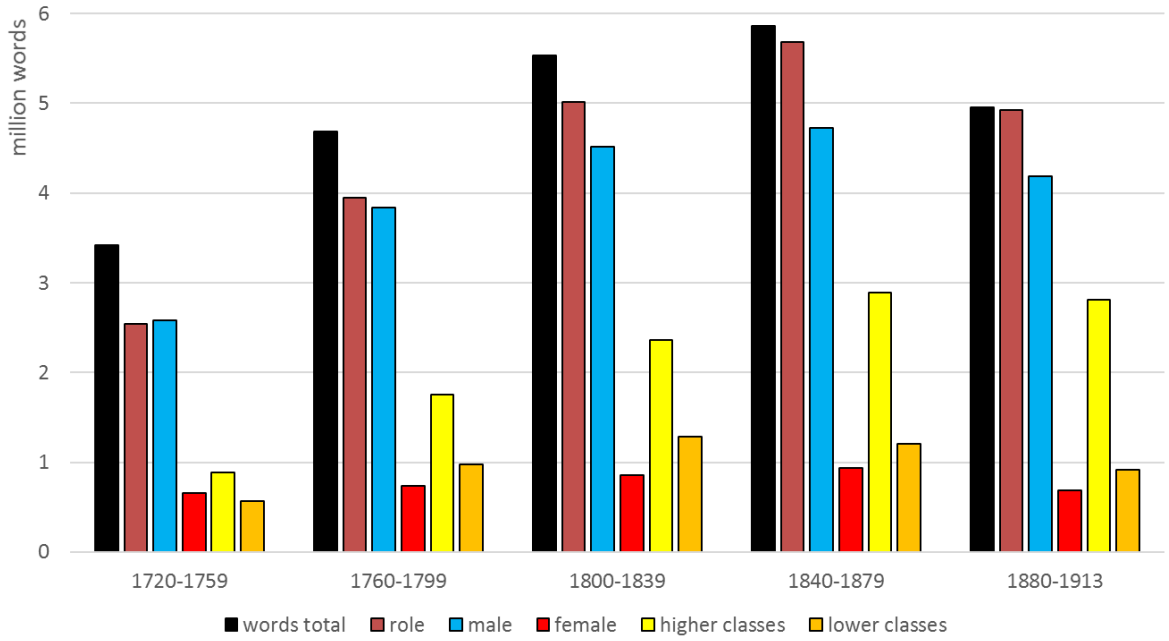

Diagram 1: Number of spoken words in *OBC 2.0* in 40-year periods

The known role of the speaker in the courtroom (rusty red) increases from 74% of the spoken words in the first period to 99% in the last period. The speaker gender (male = blue, female = red) is known for 97% of all words. At first sight the imbalance between female (16%) and male speakers (81%) might be surprising, but is explained by the fact that the court personnel was entirely male. Even though 16% sounds relatively low, this still amounts to 3.9 million words. The social class of the speaker (higher, i.e. non-manual professions = yellow; lower, i.e. manual professions = orange) is known for 64% of the spoken words in the corpus. More than two-thirds of these (44%) are from the higher social classes, which again can be explained by the fact that this includes the judges and lawyers, who contributed extensively to every trial.

5. Tagging Conventions and Procedures

OBC 2.0 consists of 637 files. The files were tagged automatically and semi-automatically with custom software.

The XML markup of the digitalized *Proceedings* was retained unchanged, with the exception of <hi> tags in utterances, which were removed whenever they occurred within a word. The following tags were added during the compilation of *OBC 2.0*: A short header indicating the date of creation of the file, a link to the *OBC 2.0* Landing Page, as well as the name of the corpus:

<text created="2016-05-17" url="http://hdl.handle.net/11858/00-246C-0000-0023-8CFB-2" corpus="The Old Bailey Corpus 2.0, 1720-1913">

Three other TEI tags were added during the *OBC* compilation, as seen in Table x:

Tag	Description
<activity>	contains a brief informal description of what a participant in a language interaction is doing other than speaking, if anything. In *OBC 2.0* such descriptions typically refer to evidence being produced in court or to the speaker's gestures.
<distinct>	identifies any word or phrase which is regarded as linguistically distinct, for example as archaic, technical, dialectal, non-preferred, etc., or as forming part of a sublanguage. In *OBC 2.0* this mostly marks English dialects or foreign accents, which are sometimes given in a quasi-phonetic spelling.
<u>	(utterance) contains a stretch of speech usually preceded and followed by silence or by a change of speaker. The opening tag contains all the sociobiographical, pragmatic and textual attributes associated with the particular utterance.

Table x. TEI tags inserted in the compilation of *OBC*

Each utterance was then tagged for sociobiographic, pragmatic and textual parameters. A typical annotated utterance looks as follows:

<u age="38" editor="" event="18500506-588" n="588" printer="William Tyler" publisher="George Hebert" role="defendant" scribe="James Drover Barnett, Alexander Buckler" sex="m" speaker="18500506-0505" trial="t18500506-976" year="1850" class="lower (6-13)" wc="26" p2="1817-1913" p3="1850-1913" p4="1818-1865" p5="1837-1875" p6="1850-1881" hisclass="9" hiscoCode="54010" hiscoCode2="" hiscoLabel="Domestic servant, general" hiscoLabel2="" nTrial="5"> I_PPIS1 am_VBM as_RG innocent_JJ as_CSA a_AT1 child_NN1 ;_; I_PPIS1 did_VDD not_XX know_VVI the_AT spoon_NN1 was_VBDZ there_RL ;_; if_CS I_PPIS1 had_VHD ,_, I_PPIS1 should_VM have_VHI objected_VVN to_II unlocking_VVG my_APPGE box_NN1 ._. </u>

(OBC t18500506-976)

The <u> … </u> tags (here colour-coded in blue) enclose the actual utterance. The opening utterance tag contains the sociobiographical (age, sex, class), pragmatic (role), textual (printer, publisher, scribe) and other attributes (e.g. year of utterance, time period of utterance for 2 to 6 periods, wc = word count) associated with this utterance. Since this example comes from the POS-tagged version, CLAWS tags (grey) are appended to the actual spoken words (black).

In the opening <u> tag, the attributes of hiscoCode and hiscoLabel indicate a speaker's occupation according to the Historical International Standard Classification of Occupations, HISCO, a database of thousands of historical occupations (van Leeuwen, Maas & Miles 2002). The attribute of hisclass assigns the speaker to a social class, following HISCLASS (van Leeuwen & Maas 2011), a social class scheme based on HISCO. The HISCLASS scheme converts the occupational HISCO codes into a system of 13 social classes. For most sociohistorical analyses it may be enough to reduce this system to a 2-class system (the attributes of class) with a higher class (non-manual professions, HISCLASS 1-5) and a lower class (manual professions, HISCLASS 6-13).

An occasional "?" after the names of scribes, printers and publishers indicates that this information was inferred by the compilers on the basis on circumstantial evidence, such as the scribes, printers, publishers of *Proceedings* immediately preceding or following the one in question.

6. Corrections

Spelling mistakes and scan errors

As a rule, the digitalized version of the *Proceedings* was left unchanged and spelling mistakes in the original were not corrected. However, scan errors were automatically and manually corrected. Sometimes smudged pages or indistinct print lead had led to letters being replaced by X's. Whenever this was the case, the scanned versions of the original *Proceedings* were consulted and, whenever possible, the X's were replaced by the corresponding letters. Scan errors such as "merged" words like *Isearched*, *hetook*, and *Iknew* and OCR errors such as *likewife* (instead of *likewise*) were identified and collected in a document (see Appendix 1) and then corrected in the whole corpus. For the identification of such errors, a word list of the entire corpus was spellchecked in a text editor. Items that were flagged as misspelled were scrutinized and added to the correction list of merged words. This mainly concerned "merged" personal pronouns, articles and conjunctions.

Punctuation

Starting with *Proceeding* 18161204, when Henry Buckler took over as scribe and T. Booth as printer / publisher, full stops were replaced by m-dashes. First sporadically, latter pervasively, as illustrated by the following scan

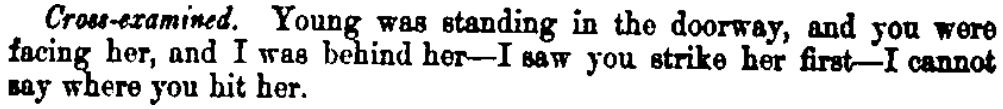
Cross-examined. Young was standing in the doorway, and you were facing her, and I was behind her—I saw you strike her first—I cannot say where you hit her.

Illustration 1. Scan from the *Proceeding* of 19 November 1888, p. 48

These m-dashes are rendered as simple hyphens in the digitalized *Proceedings*. For CLAWS to be able to recognize the strings of letters before and after such hyphens as separate words, spaces were inserted around the hyphens.

CLAWS corrections

Since CLAWS was developed for Present Day English, the CLAWS tagged files were also spot checked for wrongly assigned POS-tags:

- If a mistake was considered unique it was added to the lists of corrections that had to be implemented.
- If a mistake was considered to be part of a systematic error, the corpus was searched for other examples of this pattern and all other instances were added to the correction list. For example, CLAWS wrongly tagged sentence-final *watch-maker's* as a general adjective (*watch-maker's._JJ* instead of the correct *watch-maker_NN1 's_GE* ._.). This CLAWS-error is also found with other sentence-final s-genitives.

- Widespread CLAWS errors were replaced using regular expressions
- VVX was added to the tagset for such cases where our automatic replacements could not distinguish between a past form (VVD) and a past participle (VVN) was intended.

Detailed lists of the CLAWS corrections can be seen in Appendices 2 and 3.

Text Duplicates

While tagging the corpus, seven 7 text duplicates were discovered in the digitalized version of the *Proceedings* (they are not found in the print version). Each of the duplicated passages was deleted manually:

18680406
http://www.oldbaileyonline.org/browse.jsp?ref=18680406
search for: "on the one occasion that you saw William Desmond..."

17870110
https://www.oldbaileyonline.org/browse.jsp?ref=17870110
"Your Lordship's humanity to me, and also that of the worthy Sheriffs, from the fatal day of my conviction"

18800301
http://www.oldbaileyonline.org/browse.jsp?ref=18800301
search for: „Laird's, and the body of the letters are in”

18920307
http://www.oldbaileyonline.org/browse.jsp?ref=18920307
search for: "I have been in this trade about six years and have known of these non-genuine stamps"

18930501
http://www.oldbaileyonline.org/browse.jsp?ref=18930501
search for: „my cottage adjoins“

1900212
http://www.oldbaileyonline.org/browse.jsp?ref=19000212
search for: „I saw the advertisement on“

191110205
http://www.oldbaileyonline.org/browse.jsp?ref=19111205
search for: “I had made this confidential report for prisoner”

7. How to cite this resource

Magnus Huber, Magnus Nissel, Karin Puga (2016). Old Bailey Corpus 2.0. hdl:11858/00-246C-0000-0023-8CFB-2

8. License

References

Hitchcock, Tim, Robert Shoemaker, Clive Emsley, Sharon Howard, Jamie McLaughlin et al. 2015. /The Old Bailey Proceedings Online, 1674-1913/. www.oldbaileyonline.org, version 7.2, March 2015.

van Leeuwen, Marco H.D. and Ineke Maas. 2011. *HISCLASS: A historical international social class scheme*. Leuven: Leuven University Press.

van Leeuwen, Marco H.D., Ineke Maas & Andrew Miles. 2002. *HISCO: Historical International Standard Classification of Occupations*. Leuven University Press.

Appendix 1. Merged Words

Original	Corrections
Ihad	I had
hewas	he was
likewife	likewise
ofthe	of the
Iam	I am
Itook	I took
Iasked	I asked
assoon	as soon
hetook	he took
shewas	she was
therewas	there was
therewere	there were
thestation	the station
Idon	I don
Igave	I gave
Iheard	I heard
Iknew	I knew
Iran	I ran
Ireceived	I received
Iwent	I went
aman	a man
anythingabout	anything about
asI	as I
didnot	did not
outof	out of
prisonerwas	prisoner was
thehouse	the house
Ido	I do
Isearched	I searched
Iwould	I would
aboutthree	about three
afterthe	after the
allthe	all the
andI	and I
andhe	and he
andwent	and went
anotherman	another man
bythe	by the
donewith	done with

forthe	for the
hadbeen	had been
halfa	half a
halfan	half an
hebrought	he brought
hehad	he had
heknew	he knew
hetold	he told
himhe	him he
hispockets	his pockets
honestman	honest man
infor	in for
itin	it in
itis	it is
itit	it it
notwith	not with
ofMr	of Mr
ofthem	of them
ofthis	of this
personwho	person who
prisonerin	prisoner in
abarman	a barman
abear	a bear
ableto	able to
aboutChristmas	about Christmas
abouta	about a
aboutanother	about another
aboutfifty	about fifty
aboutfive	about five
abouthalf	about half
aboutmidnight	about midnight
aboutten	about ten
aboutthat	about that
abruised	a bruised
acarpet	a carpet
acart	a cart
AccountantGeneral	Accountant General
acheque	a cheque
aClock	a Clock
actingsergeant	acting sergeant
afalsehood	a falsehood
afemale	a female
afortnight	a fortnight
after aught	afteraught
afterhim	after him
afterlooking	after looking
afterwardssaw	afterwards saw

againtill	again till
agreat	a great
ahalf	a half
alamp	a lamp
alane	a lane
aline	a line
alump	a lump
amlost	am lost
amnot	am not
amob	a mob
amount of	amountof
amquite	am quite
andamong	and among
anddressed	and dressed
andfeet	and feet
andfound	and found
andgobling	and gobling
andher	and her
andl	and l
andlook	and look
andMcEwen	and McEwen
andmoney	and money
andmost	and most
andmy	and my
andoverhauled	and overhauled
andran	and ran
andremanded	and remanded
andreturned	and returned
andsaw	and saw
andstraw	and straw
andthat	and that
andthe	and the
andtold	and told
andwhen	and when
andwife	and wife
anend	an end
anevidence	an evidence
anirrational	an irrational
anothergentleman	another gentleman
anybodyis	anybody is
anymessage	any message
anyrobbery	any robbery
anythingto	anything to
apair	a pair
aparticular	a particular
apartner	a partner
apawnbroker	a pawnbroker

apen	a pen
apoker	a poker
appearanceof	appearance of
arrangementwas	arrangement was
asilver	a silver
askedhim	asked him
askedme	asked me
asong	a song
asure	a sure
asworking	as working
atMarlborough	at Marlborough
attemptingto	attempting to
attendingall	attending all
atthat	at that
atthem	at them
attitudethe	attitude the
authorafter	author after
authorizedyou	authorized you
AveMaria	Ave Maria
avoyage	a voyage
awareof	aware of
awarrant	a warrant
awaythe	away the
awaytwo	away two
awound	a wound
ayoung	a young
backof	back of
backsideand	backside and
becausethe	because the
bedone	be done
bedthat	bed that
BeehivePainter	Beehive Painter
beendropped	been dropped
beenin	been in
beenpawned	been pawned
beenrobbed	been robbed
beensensible	been sensible
beerit	beer it
beforeI	before I
beforethat	before that
beganjumping	began jumping
behis	be his
beingvery	being very
beready	be ready
besafe	be safe
besidesthere	besides there
besuccesful	be succesful

besworn	be sworn
betaken	be taken
bethe	be the
biscuitsfor	biscuits for
BlackHorse	Black Horse
bloodyThat	bloody That
bonesa nd	bones and
bootshe	boots he
BoroughRoad	Borough Road
bottlestands	bottle stands
boughtbook	bought book
broochfor	brooch for
BroomwellJones	Broomwell Jones
brotherconstable	brother constable
broughtthe	brought the
bruiseswere	bruises were
builtman	built man
businesshe	business he
businessthere	business there
ButolphWharf	Butolph Wharf
buttonsof	buttons of
buythem	buy them
byhis	by his
byname	by name
byword	by word
cameback	came back
camebetween	came between
cameout	came out
cameto	came to
cametowards	came towards
cameyouto	came you to
canbe	can be
cannotserve	cannot serve
capafter	cap after
CaptainWalters	Captain Walters
carriedaway	carried away
carryou	carry ou
caseof	case of
caseon	case on
causedby	caused by
certainmark	certain mark
checkapron	check apron
checkshirts	check shirts
cleanedmy	cleaned my
CliveIndia	Clive India
clockhe	clock he
ColonelBulkeley	Colonel Bulkeley

comeup	come up
comingacross	coming across
comingtowards	coming towards
commissionedofficers	commissioned officers
commitsuicide	commit suicide
committedtwo	committed two
confidencein	confidence in
consequenceof	consequence of
ConsulGeneral	Consul General
ConsulGeneral	Deputy Recorder
conversawith	conversa with
couldget	could get
couldnot	could not
counterfeitcoin	counterfeit coin
custodyat	custody at
custodyhe	custody he
custodyshe	custody she
cuthere	cut here
cutout	cut out
cuttingout	cutting out
dancein	dance in
dealmore	deal more
deceasedsitting	deceased sitting
DecemberI	December I
defendantwas	defendant was
deficientof	deficient of
DeputyRecorder	Detention Godfrey
DetentionGodfrey	Dwelling House
difficultybetween	difficulty between
diffusedthe	diffused the
directionto	direction to
distancefrom	distance from
disturbedthese	disturbed these
doesall	does all
doorappeared	door appeared
doorhe	door he
doorof	door of
dothat	do that
downa	down a
downabout	down about
downin	down in
downsenseless	down senseless
downStairs	down Stairs
draughtsno	draughts no
DwellingHouse	Fleet Bridge
dyinghe	dying he

eightpeople	eight people
eitherhim	either him
elseaccording	else according
elsenext	else next
emptyTruscott	empty Truscott
endeavouringto	endeavouring to
engagedin	engaged in
engagethose	engage those
EnglandI	England I
eveningafter	evening after
everknow	ever know
eversince	ever since
everytruss	every truss
evidencehe	evidence he
examinationof	examination of
executedthesearch	executed the search
fellowapprentice	fellow apprentice
fellowdefendant	fellow defendant
fellowprisoner	fellow prisoner
fellowservant	fellow servant
fellowservants	fellow servants
feloniouslyreceiving	feloniously receiving
findnothing	find nothing
firstbe	first be
firstlamp	first lamp
firststopped	first stopped
firstvessel	first vessel
FleetBridge	Graham Campbell
floorhe	floor he
footagainst	foot against
fordrink	for drink
forembezzling	for embezzling
forfear	for fear
forhaving	for having
forhim	for him
fourcravats	four cravats
fouryards	four yards
fouryears	four years
froma	from a
fromhis	from his
frommy	from my
fromthe	from the
fromthehen-roost	from the hen-roost
gaveher	gave her
gavehim	gave him

gaveManners	gave Manners
gavethe	gave the
getaway	get away
getdrunk	get drunk
giventhe	given the
goingdown	going down
gonethe	gone the
goodcharacter	good character
gota	got a
goto	go to
gotto	got to
gotup	got up
GrahamCampbell	John Bailey
greatdischarge	great discharge
hadany	had any
hadbetter	had better
haddone	had done
hadher	had her
hadjust	had just
hadleft	had left
hadnever	had never
hadnot	had not
halfpencein	halfpence in
hallit	hall it
handkerchieffrom	handkerchief from
hasto	has to
havebroken	have broken
havehanded	have handed
havesaid	have said
haveserved	have served
haveyouseen	have you seen
havingbeen	having been
headmitted	he admitted
heartto	heart to
hebought	he bought
hedelivered	he delivered
hefinished	he finished
hegave	he gave
heintroduced	he introduced
heknocked	he knocked
heleft	he left
helent	he lent
heopened	he opened
hepicked	he picked
heran	he ran
herdaughter	her daughter
herdesk	her desk

herhouse	her house
hermother	her mother
heror	her or
hershe	her she
hesaw	he saw
heseemed	he seemed
hestated	he stated
hethen	he then
hewould	he would
higherup	higher up
highlyrespectable	highly respectable
himand	him and
himchange	him change
himin	him in
himinto	him into
himlend	him lend
himlying	him lying
himto	him to
himwith	him with
hisabsence	his absence
hiscoat	his coat
hisdelivering	his delivering
hisfather	his father
hisforehead	his forehead
hishand	his hand
hishead	his head
hisperson	his person
hispocket	his pocket
histable	his table
histrial	his trial
homefrom	home from
homething	something
hotbegun	not begun
hotwater	hot water
hourof	hour of
househe	house he
househeard	house heard
howhe	how he
howmuch	how much
howto	how to
Ibeg	I beg
Ibelieve	I believe
Iblew	I blew
Ibought	I bought
Icannot	I cannot
Ichanged	I changed
Icould	I could

Icried	I cried
Iendeavoured	I endeavoured
Iexamined	I examined
ifhe	if he
ifour	if our
Ihere	I here
Iimmediately	I immediately
Iintended	I intended
Ikept	I kept
illfeeling	ill feeling
illtemper	ill temper
Ilost	I lost
Imade	I made
Imet	I met
Imight	I might
Imissed	I missed
importancepassed	importance passed
inabout	in about
incase	in case
incharge	in charge
incustody	in custody
inexperiencedthey	inexperienced they
inmy	in my
Inoticed	I noticed
inPenal	in Penal
inquestion	in question
inShepherd	in Shepherd
insideincluding	inside including
intoMrs	into Mrs
Iproduce	I produce
Irecollect	I recollect
Iremember	I remember
ironsquietl y\.	irons quietly.
isa	is a
islabouring	is labouring
ismine	is mine
Ispoke	I spoke
Istarted	I started
Istood	I stood
isunder	is under
isworth	is worth
itfor	it for
ithad	it had
ithappened	it happened
Ithink	I think
itnow	it now
Itold	I told

iton	it on
itsfeatures	its features
itwhen	it when
Iunlocked	I unlocked
Iwant	I want
Iwill	I will
JohnBailey	John Hodgkinson
JohnHodgkinson	King William
justas	just as
KingWilliam	Love Lane
knewthe	knew the
knownothing	know nothing
knowthis	know this
largestrevolver	largest revolver
latehe	late he
laterin	later in
lawsuitabout	lawsuit about
leaveNo	leave No
leftmeto	left me to
leftthe	left the
lengthfor	length for
letcarriers	let carriers
lethim	let him
lethis	let his
lightin	light in
likeit	like it
liquorwhen	liquor when
littlego	little go
livedhe	lived he
longtime	long time
lookedat	looked at
lookedout	looked out
LoveLane	Millie Marsh
lyingto	lying to
ma hogany	mahogany
madeshare	made share
managainst	man against
manbrought	man brought
manyhave	many have
markthe	mark the
marriedto	married to
materialsto	materials to
mebefore	me before
medown	me down
mercyby	mercy by
merelyspoke	merely spoke
methat	me that

militaryNavalDuty	military Naval Duty
MillieMarsh	Miss Storey
MissStorey	Peter Carthew
moneyfrom	money from
monthsago	months ago
mostrespectable	most respectable
mouththe	mouth the
movefrom	move from
mustin	must in
myaccomplice	my accomplice
myback	my back
myfather	my father
myfellow	my fellow
myhead	my head
myknowledge	my knowledge
mylittle	my little
mymind	my mind
mymother	my mother
mymy	my my
myname	my name
myostler	my ostler
myquestions	my questions
mystatement	my statement
mywife	my wife
myworkshop	my workshop
namein	name in
neverconsulted	never consulted
nextday	next day
nightthe	night the
nightwhen	night when
ninecreditors	nine creditors
noclock	no clock
noconsequence	no consequence
noHarm	no Harm
nomarks	no marks
notany	not any
notappear	not appear
notaware	not aware
notentitled	not entitled
notexamined	not examined
notfind	not find
nothingat	nothing at
nothingextraordinary	nothing extraordinary
nothingwas	nothing was
noticedthem	noticed them
notlarge	not large

notpay	not pay
notpromise	not promise
notremember	not remember
notsay	not say
notseeany	not see any
nottake	not take
ofany	of any
ofascertaining	of ascertaining
ofAugust	of August
ofbricks	of bricks
ofBrighton	of Brighton
ofcloth	of cloth
ofcotton	of cotton
ofcourse	of course
ofDonnelly	of Donnelly
offersabond	offers a bond
offhe	off he
officerfrom	officer from
ofgin	of gin
ofhair	of hair
ofhaving	of having
ofJune	of June
oflinen	of linen
ofLittle	of Little
ofMarch	of March
ofserum	of serum
ofThompson	of Thompson
oldit	old it
onand	on and
onbehalf	on behalf
onconsideration	on consideration
oneeighth	one eighth
onefourth	one fourth
oneof	one of
oneon	one on
onepound	one pound
onewilling	one willing
ongetting	on getting
onhis	on his
onin	on in
onthis	on this
onwe	on we
onwhen	on when
opensabout	opens about
oppositeGravel	opposite Gravel
orthree	or three
ortwenty	or twenty

othermen	other men
othermoneys	other moneys
otherof	other of
otherside	other side
othertwo	other two
otherwiseCONSTANT	otherwise CONSTANT
oursolicitors	our solicitors
ourtide	our tide
outerdoor	outer door
outfittingshop	outfitting shop
outhis	out his
outin	out in
outsidethe	outside the
overbefore	over before
overher	over her
overon	over on
owndresses	own dresses
ownhandkerchief	own handkerchief
particularabout	particular about
pavementhe	pavement he
payingsome	paying some
PeterCarthew	Royal Exchange
placeit	place it
positivelyswear	positively swear
prisonerbegged	prisoner begged
prisonerforgetting	prisoner forgetting
prisonergot	prisoner got
prisonerJames	prisoner James
prisonerlying	prisoner lying
prisonermany	prisoner many
prisonersaid	prisoner said
prisonerstruck	prisoner struck
prisonerswere	prisoners were
producebefore	produce before
producedare	produced are
producedto	produced to
promissorynote	promissory note
prosecutorrolled	prosecutor rolled
prosecutorsaid	prosecutor said
prosecutrixwith	prosecutrix with
pulledout	pulled out
purposesof	purposes of
ranafter	ran after
ranaway	ran away
rantowards	ran towards
rappinga	rapping

rathersuspiciously	rather suspiciously
rattlethat	rattle that
readilysucked	readily sucked
receivedthese	received these
receivingward	receiving-ward
recollectthe	recollect the
relatingto	relating to
relievingoverseer	relieving overseer
representationto	representation to
requestedme	requested me
returnsare	returns are
riggingall	rigging all
RoyalExchange	Royal Oak
RoyalOak	Silk Handkerchief
saidhe	said he
saidI	said I
saidthat	said that
saidWho	said Who
sameday	same day
sameevening	same evening
sawone	saw one
sawthem	saw them
sayingit	saying it
saysI	says I
saysmy	says my
saysshe	says she
saywhat	say what
seemedto	seemed to
seensingle	seen single
selfsee	self see
sellit	sell it
sendin	send in
sendingearlier	sending earlier
sentit	sent it
servedthree	served three
sevenlinen	seven linen
sevenshilling	seven shilling
sewingsilk	sewing silk
shawlaside	shawl aside
sheasked	she asked
shecame	she came
shecomplained	she complained
shefollowed	she followed
shegave	she gave
shehad	she had
shekept	she kept
sheleant	she leant

sheonly	she only
shesuffered	she suffered
shetook	she took
shewent	she went
shillingthat	shilling that
shockwhen	shock when
shortlyafter	shortly after
SilkHandkerchief	Sir William
silverand	silver and
sincepaid	since paid
sincethis	since this
singleman	single man
singlewoman	single woman
SirWilliam	Smith was
sistershe	sister she
sixmonths	six months
slipdown	slip down
Smithwas	Smithwas
sodecomposed	so decomposed
sodisguised	so disguised
soldierat	soldier at
soldit	sold it
someof	some of
somestolen	some stolen
somethings	some things
somethingthat	something that
sometimesthere	sometimes there
sortout	sort out
sovereignwas	sovereign was
spoketo	spoke to
stablewe	stable we
Standingon	Standing on
stateyouhad	state you had
stationand	station and
StreetCheapside	Street Cheapside
streethe	street he
struckagainst	struck against
struckChaplen	struck Chaplen
submittedthere	submitted there
SuperbeMan	Superbe Man
takethem	take them
takeup	take up
takinghim	taking him
takingmy	taking my
Talkendon	Talkend on
Taplinwere	Taplin were
tea-potand	tea-pot and

testimonialyou	testimonial you
thathe	that he
thathere	that here
thathouse	that house
thatman	that man
thatMartyn	that Martyn
thatreflects	that reflects
thattook	that took
thatused	that used
thatwatch	that watch
thatyou	that you
theaccident	the accident
theagreement	the agreement
thearticles	the articles
theattorney	the attorney
theauction	the auction
theauthority	the authority
thebaby	the baby
thebags	the bags
thebaskets	the baskets
theblind	the blind
theblow	the blow
thebone	the bone
theBritish	the British
thebruise	the bruise
thebullet	the bullet
thebundle	the bundle
thecase	the case
theconstable	the constable
thecopper	the copper
thecorner	the corner
thecushion	the cushion
thedecesed	the decesed
thedoor	the door
theenvelope	the envelope
theevidence	the evidence
theeyebrow	the eyebrow
thefather	the father
thefloor	the floor
thegangway	the gangway
thegate	the gate
theGoods	the Goods
thehalf	the half
thehorses	the horses
thehospital	the hospital
theinitials	the initials
thelady	the lady

thelandlord	the landlord
thelast	the last
theman	the man
themfinancial	them financial
themif	them if
themplay	them play
thenand	then and
thenew	the new
thenremanded	then remanded
thenumber	the number
theoffice	the office
theofficer	the officer
theone	the one
theorder	the order
theordinary	the ordinary
thePerson	the Person
Thepoliceman	The policeman
thePortland	the Portland
thepremises	the premises
theprisoners	the prisoners
thereunder	there under
theroom	the room
therope	the rope
thesale	the sale
theSaturday	the Saturday
thesaw	the saw
theScotch	the Scotch
theseare	these are
theSecondary	the Secondary
theseto	these to
theship	the ship
theSociety	the Society
thestain	the stain
thestreet	the street
thetickets	the tickets
thetime	the time
thetrade	the trade
theviolin	the violin
thewaggon	the waggon
thewarrant	the warrant
thewatch	the watch
thewater	the water
thewoman	the woman
theyappear	they appear
theyasked	they asked
theyboth	they both
Theyhave	They have

theywent	they went
thighit	thigh it
thingsof	things of
thingsout	things out
thingsthat	things that
thinkyou	think you
thirteenyears	thirteen years
thisagreement	this agreement
thisknife	this knife
thisnotice	this notice
ThisNow	This Now
thisprospectus	this prospectus
thissteel	this steel
Thiswas	This was
threeminutes	three minutes
Threeor	Three or
threetowels	three towels
throwmy	throw my
timehe	time he
timethat	time that
toas	to as
toascertain	to ascertain
toattend	to attend
toCrevey	to Crevey
toEngland	to England
togive	to give
togo	to go
toher	to her
tohim	to him
toldher	told her
toldthe	told the
tomercy	to mercy
toMr	to Mr
ToMrs	To Mrs
toMrs	to Mrs
tookthem	took them
toParker	to Parker
torequest	to request
tosomething	to something
totake	to take
toyou	to you
trousersand	trousers and
Tuesdaymorning	Tuesday morning
turnedround	turned round
twoother	two other
twopair	two pair
twothirds	two thirds

understoodhim	understood him
undersuspicious	under suspicious
underthe	under the
upin	up in
verystrange	very strange
veryweak	very weak
verywell	very well
wagonhe	wagon he
Waltoncame	Walton came
wantedthe	wanted the
wardspicked	wards picked
warmgin	warm gin
wasbad	was bad
wascalled	was called
wasexamined	was examined
wasfrom	was from
wasin	was in
wasit	was it
wasnottheoalytime	was not the only time
wasOne	was One
wasout	was out
waspaid	was paid
waspassing	was passing
waspresent	was present
wasput	was put
wassome	was some
wasthen	was then
wastime	was time
wasto	was to
wastook	was took
wastrue	was true
wasworth	was worth
watchingunder	watching under
wayas	way as
weekbefore	week before
weekfor	week for
weeksbefore	weeks before
wegot	we got
wehad	we had
wehanded	we handed
Wehereby	We hereby
wentaft	went aft
wentaway	went away
wentback	went back
wentin	went in
wenton	went on

wentto	went to
werealso	were also
weregoing	were going
werein	were in
werenot	were not
wereout	were out
weresent	were sent
werethere	were there
weretwelve	were twelve
wereutensils	were utensils
wesaw	we saw
Weshall	We shall
weshould	we should
Whatman	What man
whatnow	what now
whenthe	when the
whenyou	when you
Whereis	Where is
wheresthe	wheres the
whetherhe	whether he
whetherMr	whether Mr
WhiteHart	White Hart
WhiteHorse	White Horse
Whoareyou	Who are you
wholeof	whole of
whoput	who put
wildrabbits	wild rabbits
Winethen	Wine then
withanother	with another
withher	with her
withhim	with him
withme	with me
withOliver	with Oliver
withoutAldgate	without Aldgate
withstealing	with stealing
withyou	with you
witnessthe	witness the
workfor	work for
worstedcord	worsted cord
wouldbe	would be
wouldhave	would have
wouldsoon	would soon
woundunder	wound under
wrappedin	wrapped in
writedown	write down
writtendefence	written defence
yardsfrom	yards from

yardshe	yards he
Yearin	Year in
yearshe	year she
youany	you any
youlost	you lost
yourbrother	your brother
yourchild	your child
yourthreat	your threat
yousee	you see
youwere	you were

Appendix 2. CLAWS Corrections

Mistakes	Corrections
Stop_NN1 Thief_NN1!_!	Stop_VV0 Thief_NN1 !_!
Stop_NN1 Thief_NN1	Stop_VV0 Thief_NN1
Stop_NN1 Thieves_NN2	Stop_VV0 Thieves_NN2
Stop_NN1 theif_NN1	Stop_VV0 theif_NN1
Stop_NN1 thief_NN1!_!	Stop_VV0 thief_NN1 !_!
Stop_NN1 thief_NN1	Stop_VV0 thief_NN1
stop_NN1 Duke_NN1	stop_VV0 Duke_NP1
stop_NN1 Thief_NN1!_!	stop_VV0 Thief_NN1 !_!
stop_NN1 Thief_NN1	stop_VV0 Thief_NN1
stop_NN1 pickpocket_NN1	stop_VV0 pickpocket_NN1
stop_NN1 thief_NN1!_!	stop_VV0 thief_NN1 !_!
stop_NN1 thief_NN1	stop_VV0 thief_NN1
stop_NN1 thieves_NN2	stop_VV0 thieves_NN2
depo_NN1 s_ZZ1 'd_VM	depos'd_VVD
mis_NN1 s_ZZ1 'd_VM	miss'd_VVN
o'Mornings_NN2	o'_II Mornings_NN2
o'Night_NN1	o'_II Night_NN1
o'Nights_NN2	o'_II Nights_NN2
o'Nights._NNU	o'_II Nights_NN2 ._.
o'Nights_VVZ	o'_II Nights_NN2
o'Saturday_NN1	o'_II Saturday_NPD1
o'score_NN1	o_II score_NN1
o'Ship_NN1	o'_II Ship_NN1
o'ship_NN1	o_II ship_NN1
o'th_NN1 Head_NN1	o'_II th'_AT1 Head_NN1
o'the_NN1 Clock_NN	o'the_RA21 Clock_RA22
o'the_NN1 Face_NN1	o'_II th'_AT1 Face_NN1
o'top_NN1	o'_II top_NN1
o'Window_VVI	o'_IO Window_NN1
being_VBG aiding_NN1	being_VBG aiding_VVG
being_VBG going_JJ	being_VBG going_VVG
being_VBG eating_NN1	being_VBG eating_VVG
being_VBG hunting_NN1	being_VBG hunting_VVG
being_VBG hurrying_JJ	being_VBG hurrying_VVG
being_VBG kneeling_JJ	being_VBG kneeling_VVG
being_VBG lighting_NN1	being_VBG lighting_VVG
being_VBG repairing_JJ	being_VBG repairing_VVG
being_VBG selling_NN1	being_VBG selling_VVG
being_VBG standing_NN1	being_VBG standing_VVG
being_VBG struggling_JJ	being_VBG struggling_VVG
being_VBG talking_JJ	being_VBG talking_VVG
being_VBG walking_NN1	being_VBG walking_VVG

amongest_JJT	amongest_II
breakfest_JJT	breakfest_NN1
Confest_JJT	Confest_VVD
Confest_NP1	Confest_VVD
has_VHZ confest_JJT	has_VHZ confest_VVN
had_VHD confest_JJT	had_VHD confest_VVN
having_VHG confest_JJT	having_VHG confest_VVN
confest_JJT	confest_VVD
I_MC1 O_ZZ1 U_ZZ221 's_ZZ222	IOU's_NN2
'I_UH O_UH U'_NP1	IOU'_NN1
I_ZZ1 O_ZZ1 U_JJ	IOU_NN1
<lc> IO_NP1 </lc> U_JJ	<lc>IOU_NN1</lc>
I_ZZ1 O_ZZ1 U_JJ	IOU_NN1
10_MC U_JJ	IOU_NN1
never_RR receipt_NN1	never_RR receipt_VV0
always_RR receipt_NN1	always_RR receipt_VV0
please_RR receipt_NN1	please_RR receipt_VV0
did_VDD not_XX receipt_NN1	did_VDD not_XX receipt_VVI
slipt_NN1	slipt_VVD
reotipt_NN1	receipt_NN1
were_VBDR tript_NN1 up_RP	were_VBDR tript_VVN up_RP
suddenly_RR tript_NN1 up_RP	suddenly_RR tript_VVN up_RP
takin_VVG g_ZZ1	taking_VVG
having_VHG stept_NN1	having_VHG stept_VVN
stept_VV0	stept_VVD
stept_VVI	stept_VVD
stept_NN1	stept_VVD
to_II sware_NN1	to_TO sware_VVI
_XX sware_NN1	_XX sware_VVI
_VM sware_NN1	_VM sware_VVI
sware_NN1	sware_VV0
witnessmy_JJ	witness_NN1 my_AAPGE
was_VBDZ ript_NN1	was_VBDZ ript_VVN
ript_NN1	ript_VVD
you_PPY 'd_VM 'a_UH blest_JJ	you_PPY 'd_VM 'a_VHI blest_VVN
going_VVG 'o_UH plead_VV0	going_VVGK 'o_TO plead_VVI
o_UH clock_NN1	o_RA21 clock_RA22
'tis_JJ	't_PPH1 is_VB0
'tis_NN1	't_PPH1 is_VB0
'a_UH	'a_AT1
'A_UH	'A_AT1

'o_UH him_PPHO	'o_IO him_PPHO
val_NN1 ._.	val._NN1
bank_NN1 note_VV0	bank_NN1 note_NN1
Bank_NN1 note_VV0	Bank_NN1 note_NN1
Bank_NN1 Note_VV0	Bank_NN1 Note_NN1
bank_NN1 notes_VVZ	bank_NN1 notes_NN2
Bank_NN1 notes_VVZ	Bank_NN1 notes_NN2
Bank_NN1 Notes_VVZ	Bank_NN1 Notes_NN2
1/2lbs_FU	1/2_MC lbs_NNU2
there_EX were_VBDR houses_NN2 building_NN1	there_EX were_VBDR houses_NN2 building_VVG
incident_NN1 to_II	incident_JJ to_II
pounds_NN2	pounds_NNU2
Pounds_NN2	Pounds_NNU2
pound_NN1	pound_NNU
Pound_NN1	Pound_NNU
d'ye_NN1	d'_VD0 ye_PPY
d'ye_VV0	d'_VD0 ye_PPY
a_AT1 Dye_NP1	a_AT1 Dye_NN1
blackest_JJT Dye_NP1	blackest_JJT Dye_NN1
Bitch_NN1 Dye_NP1	Bitch_NN1 Dye_VV0
to_II dye_NN1	to_TO dye_VVI
to_TO dye_NN1	to_TO dye_VVI
dye_NN1 by_II Inches_NNU2	dye_VV0 by_II Inches_NNU2
_VM dye_NN1	_VM dye_VVI
_XX dye_NN1	_XX dye_VVI
a_AT1 Lye_NP1	a_AT1 Lye_NN1
a_AT1 damned_JJ Lye_NP1	a_AT1 damned_JJ Lye_NN1
to_II Lye_NP1	to_TO Lye_VVI
any_DD Lye_NP1	any_DD Lye_NN1
I_MC1	I_PPIS1
settest_JJT	settest_VV0
imprest_NN1	imprest_VVN
forwardest_JJT	forwardest_RL
mean-drest_JJT	mean-drest_JJ
prest_JJT him_PPHO	prest_VVD him_PPHO
they_PPHS2 prest_JJT	they_PPHS2 prest_VVD
prest_JJT	prest_VVN
prest._NNU	prest_VVN ._.
being_NN1 undrest_JJT	being_VBG undrest_JJ
being_VBG undrest_JJT	being_VBG undrest_JJ
was_VBDZ undrest_JJT	was_VBDZ undrest_JJ
or_CC undrest_JJT	or_CC undrest_JJ
partly_RR undrest_JJT	partly_RR undrest_JJ

undrest_JJT	undrest_VVD
Drest_JJT	Drest_JJ
Drest_NP1	Drest_JJ
drest_JJT	drest_VVN
Faulcon-bridge's._JJ	Faulcon-bridge_NP1 's_GE ._.
Pawn-broker's._JJ	Pawn-broker_NN1 's_GE ._.
brother-in-law's._JJ	brother-in-law_NN1 's_GE ._.
coach-maker's._JJ	coach-maker_NN1 's_GE ._.
copper-smith's._JJ	copper-smith_NN1 's_GE ._.
corn-chandler's._JJ	corn-chandler_NN1 's_GE ._.
fortune-teller's._JJ	fortune-teller_NN1 's_GE ._.
grand-mother's._JJ	grand-mother_NN1 's_GE ._.
green-grocer's._JJ	green-grocer_NN1 's_GE ._.
mantua-maker's._JJ	mantua-maker_NN1 's_GE ._.
pastry-cook's._JJ	pastry-cook_NN1 's_GE ._.
pawn-broker's._JJ	pawn-broker_NN1 's_GE ._.
watch-maker's._JJ	watch-maker_NN1 's_GE ._.
Silver-spinner's._JJ	Silver-spinner_NN1 's_GE ._.
'till_VV0	'till_CS
'till_NN1	'till_CS
'til_VV0	'til_CS
Till_NN1	'Till_CS
'Till_VV0	'Till_CS
'till_VVI	'till_CS
"_JJ	"_"
"_NN1	"_"
"_NP1	"_"
"_VV0	"_"
"_VVI	"_"
was_58 '_GE	was_VBDZ
says_58 '_GE	says_VVZ
ma_06 'm_VBM	ma'm_NN1
1_MC11	1_MC1
thro'out_VV0	thro'out_II
thro'a_NN1	thro'_II a_AT1
Drivethro'_NP1	Drive_VV0 thro'_II
thro'._NNU	thro'_II ._.
thro'_JJ	thro'_II
thro'_NN1	thro'_II
thro'_VVI	thro'_II
thro'_VV0	thro'_II
`_" em_FU	'em_PPHO2
bank_NN1 notes_VVZ	bank_NN1 notes_NN2
"_JJ	"_"
"_NN1	"_"

"_NP1	"_"
"_VV0	"_"
"_VVI	"_"
was_58 '_GE	was_VBDZ
says_58 '_GE	says_VVZ
ma_06 'm_VBM	ma'm_NN1
Â_NULL £_NULL	£_NNU
£_NULL	£_NNU
([a-zA-Z-]+)_VVZ \'d_[A-Z]+	\1'd_VVX
([a-zA-Z-]+)_VVD \'d_[A-Z]+	\1'd_VVX
([a-zA-Z-]+)_FU \'d_[A-Z]+	\1'd_VVX
([a-zA-Z-]+)_NN \'d_[A-Z]+	\1'd_VVX
([a-zA-Z-]+)_RP \'d_[A-Z]+	\1'd_VVX

Appendix 3. CLAWS Corrections Punctuation

Mistakes	Corrections
me._NNU	me_PPIO1 ._.
in._NNU	in_RP ._.
Co._NP1	Co._NN1
us._NNU	us_PPIO2 ._.
to._NNU	to_II ._.
is._NNU	is_VBZ ._.
court._NNU	court_NN1 ._.
Watch-house._JJ	Watch-house_NN1 ._.
public-house._JJ	public-house_NN1 ._.
prosecutor._NNU	prosecutor_NN1 ._.
ill._NNU	ill_JJ ._.
Hand-writing._JJ	Hand-writing_NN1 ._.
it._NNU	it_PPH1 ._.
half-crown._JJ	half-crown_NN1 ._.
St._NNL1	St._NP1
Newgate._NP1	Newgate_NP1 ._.
prosecutrix._NNU	prosecutrix_NN1 ._.
them._NNU	them_PPHO2 ._.
produced._NNU	produced_VVX ._.
breeches._NNU	breeches_NN2 ._.
station-house._JJ	station-house_NN1 ._.
Compter._NP1	Compter_NN1 ._.
cloaths._NNU	cloaths_NN2 ._.
halfpence._NNU	halfpence_NN1 ._.
police-station._JJ	police-station_NN1 ._.
hand._NNU	hand_NN1 ._.
pawnbroker._NNU	pawnbroker_NN1 ._.
trowsers._NNU	trowsers_NN2 ._.
Six-pence._JJ	Six-pence_NN1 ._.
in._II	in_RP ._.
intoxicated._NNU	intoxicated_JJ ._.
spoons._NNU	spoons_NN2 ._.
coals._NNU	coals_NN2 ._.
pawned._NNU	pawned_VVX ._.
buckles._NNU	buckles_NN2 ._.
tap-room._JJ	tap-room_NN1 ._.
Sessions._NP1	Sessions_NN2 ._.
Round-house._JJ	Round-house_NN1 ._.
sovereigns._NNU	sovereigns_NN2 ._.
post-office._JJ	post-office_NN1 ._.
coppers._NNU	coppers_NN2 ._.
bed-room._JJ	bed-room_NN1 ._.

pawnbrokers._NNU	pawnbrokers_NN2 ._.
Cheapside._NP1	Cheapside_NP1 ._.
Lambeth._NP1	Lambeth_NP1 ._.
public-house._NNU	public-house_NN1 ._.
watchman._NNU	watchman_NN1 ._.
Cloaths._NP1	Cloaths_NN2 ._.
St._NNB	St._NP1
felony._NNU	felony_NN1 ._.
insensible._NNU	insensible_JJ ._.
Pimlico._NP1	Pimlico_NP1 ._.
happen'd._NNU	happen'd_VVX ._.
him._NNU	him_PPHO1 ._.
who._NNU	who_PNQS ._.
sixpences._NNU	sixpences_NN2 ._.
Hoxton._NP1	Hoxton_NP1 ._.
Houndsditch._NP1	Houndsditch_NP1 ._.
carman._NNU	carman_NN1 ._.
Welch._NP1	Welch_NP1 ._.
read._NNU	read_VVX ._.
Aldgate._NP1	Aldgate_NP1 ._.
V._NNU	V._MC
half-sovereign._JJ	half-sovereign_NN1 ._.
counting-house._JJ	counting-house_NN1 ._.
Breeches._NP1	Breeches_NN2 ._.
one._NNU	one_PN1 ._.
Buckles._NP1	Buckles_NN2 ._.
depos'd._NNU	depos'd_VVX ._.
halfpenny._NNU	halfpenny_NN1 ._.
garret._NNU	garret_NN1 ._.
lodgers._NNU	lodgers_NN2 ._.
coachman._NNU	coachman_NN1 ._.
tankard._NNU	tankard_NN1 ._.
W._NNU	W._NP1
Millan._NP1	Millan_NP1 ._.
Spitalfields._NP1	Spitalfields_NP1 ._.
V._II	V._MC
absconded._NNU	absconded_VVX ._.
dishonoured._NNU	dishonoured_JJ ._.
mare._NNU	mare_NN1 ._.
sessions._NNU	sessions_NN2 ._.
Court._NP1	Court_NN1 ._.
a-piece._NNU	a-piece_RA ._.
chaise._NNU	chaise_NN1 ._.
don't._NNU	do_VM n't_XX ._.
Shadwell._NP1	Shadwell_NP1 ._.

lodging-house._JJ	lodging-house_NN1 ._.
property._NNU	property_NN1 ._.
robb'd._NNU	robb'd_VVX ._.
Rotherhithe._NP1	Rotherhithe_NP1 ._.
cash-book._JJ	cash-book_NN1 ._.
twelvemonth._NNU	twelvemonth_NN1 ._.
pocket-book._JJ	pocket-book_NN1 ._.
Lord-Mayor._NP1	Lord-Mayor_NN1 ._.
barman._NNU	barman_NN1 ._.
bedstead._NNU	bedstead_NN1 ._.
hang'd._NNU	hang'd_VVX ._.
Mansion-house._JJ	Mansion-house_NN1 ._.
cask._NNU	cask_NN1 ._.
Co._FO	Co._NN1
half-crowns._NNU	half-crowns_NN2 ._.
taproom._NNU	taproom_NN1 ._.
Felony._NP1	Felony_NN1 ._.
compter._NNU	compter_NN1 ._.
half-crowns._JJ	half-crowns_NN2 ._.
deposition._NNU	deposition_NN1 ._.
bricklayer._NNU	bricklayer_NN1 ._.
forenoon._NNU	forenoon_NN1 ._.
alehouse._NNU	alehouse_NN1 ._.
victuals._NNU	victuals_NN2 ._.
M._NNO	M._NP1
cashed._NNU	cashed_VVX ._.
Half-pence._JJ	Half-pence_NN1 ._.
forgeries._NNU	forgeries_NN2 ._.
Produced._NP1	Produced_VVX ._.
lodger._NNU	lodger_NN1 ._.
indictment._NNU	indictment_NN1 ._.
chissel._NNU	chissel_NN1 ._.
dresser._NNU	dresser_NN1 ._.
pony._NNU	pony_NN1 ._.
dwelling-house._JJ	dwelling-house_NN1 ._.
cash-box._JJ	cash-box_NN1 ._.
cloths._NNU	cloths_NN2 ._.
C._ZZ1	C_NN1 ._.
met._NNU	met_VVX ._.
fowls._NNU	fowls_NN2 ._.
wt._NN1	wt._NNU
prisoner._NNU	prisoner_NN1 ._.
farthings._NNU	farthings_NN2 ._.
lighterman._NNU	lighterman_NN1 ._.
horseback._NNU	horseback_NN1 ._.

Drury-lane._NNU	Drury-lane_NP1 ._.
thoroughfare._NNU	thoroughfare_NN1 ._.
me._PPIO1	me_PPIO1 ._.
Hearne._NP1	Hearne_NP1 ._.
W._ND1	W._NP1
kill'd._NNU	kill'd_VVX ._.
half-a-crown._JJ	half-a-crown_NN1 ._.
mention'd._NNU	mention'd_VVX ._.
Covent-Garden._NP1	Covent-Garden_NP1 ._.
florins._NNU	florins_NN2 ._.
Alehouse._NP1	Alehouse_NN1 ._.
Dials._NP1	Dials_NN2 ._.
Moorfields._NP1	Moorfields_NP1 ._.
rob'd._NNU	rob'd_VVX ._.
Victuals._NP1	Victuals_NN2 ._.
se'nnight._NNU	se'nnight_NN1 ._.
Fellow-_NN1	Fellow-_NN1
character._NNU	character_NN1 ._.
portmanteau._NNU	portmanteau_NN1 ._.
name._NNU	name_NN1 ._.
scuffle._NNU	scuffle_NN1 ._.
Highgate._NP1	Highgate_NP1 ._.
bobbins._NNU	bobbins_NN2 ._.
Aldermanbury._NP1	Aldermanbury_NP1 ._.
Holbourn._NP1	Holbourn_NP1 ._.
workman._NNU	workman_NN1 ._.
p._NNU	p._NN1
Bed-side._JJ	Bed-side_NN1 ._.
Drury-Lane._NP1	Drury-Lane_NP1 ._.
liv'd._NNU	liv'd_VVX ._.
Perreau._NP1	Perreau_NP1 ._.
on't._NNU	on_II 't_PPH1 ._.
bushels._NNU	bushels_NN2 ._.
Oxford-street._NNU	Oxford-street_NP1 ._.
Machattie._NP1	Machattie_NP1 ._.
affrighted._NNU	affrighted_JJ ._.
moneys._NNU	moneys_NN2 ._.
Rohan._NP1	Rohan_NP1 ._.
lanthorn._NNU	lanthorn_NN1 ._.
street-door._JJ	street-door_NN1 ._.
life-time._JJ	life-time_NN1 ._.
half-crown._NNU	half-crown_NN1 ._.
forgiveness._NNU	forgiveness_NN1 ._.
Longford._NP1	Longford_NP1 ._.
frock._NNU	frock_NN1 ._.

e._ZZ1	e_NN1 ._.
Chaise._NP1	Chaise_NN1 ._.
Round-House._NP1	Round-House_NN1 ._.
Hay-market._JJ	Hay-market_NP1 ._.
locket._NNU	locket_NN1 ._.
beds._NNU	beds_NN2 ._.
candlestick._NNU	candlestick_NN1 ._.
Tripland._NP1	Tripland_NP1 ._.
melted._NNU	melted_NVVX ._.
day-book._JJ	day-book_NN1 ._.
day-light._JJ	day-light_NN1 ._.
up-stairs._JJ	up-stairs_JJ ._.
stopp'd._NNU	stopp'd_VVX ._.
East-Smithfield._NP1	East-Smithfield_NP1 ._.
Tankard._NP1	Tankard_NN1 ._.
paper._NNU	paper_NN1 ._.
Twelvemonth._NP1	Twelvemonth_NN1 ._.
Post-office._NNU	Post-office_NN1 ._.
Rag-Fair._NP1	Rag-Fair_NP1 ._.
Mews._NP1	Mews_NP1 ._.
water-closet._JJ	water-closet_NN1 ._.
New-Prison._NP1	New-Prison_NP1 ._.
original._NNU	original_NN1 ._.
mantel-piece._JJ	mantel-piece_NN1 ._.
skittles._NNU	skittles_NN2 ._.
Custom-house._JJ	Custom-house_NN1 ._.
coining._NNU	coining_NN1 ._.
gentlewoman._NNU	gentlewoman_NN1 ._.
Odell._NP1	Odell_NP1 ._.
Blackwall._NP1	Blackwall_NP1 ._.
whiskers._NNU	whiskers_NN2 ._.
watchhouse._NNU	watchhouse_NN1 ._.
errand._NNU	errand_NN1 ._.
bowels._NNU	bowels_NN2 ._.
cabman._NNU	cabman_NN1 ._.
Finsbury._NP1	Finsbury_NP1 ._.
Publick-house._JJ	Publick-house_NN1 ._.
half-and-half._JJ	half-and-half_NN1 ._.
qrs._NNU	qrs._NN2
Judd._NP1	Judd_NP1 ._.
Chance-Medley._NP1	Chance-Medley_NP1 ._.
Poulson._NP1	Poulson_NP1 ._.
witness._NNU	witness_NN1 ._.
erysipelas._NNU	erysipelas_NN1 ._.
Rosemary-Lane._NP1	Rosemary-Lane_NP1 ._.

New-stairs._JJ	New-stairs_NP1 ._.
Ware-house._JJ	Ware-house_NN1 ._.
shavings._NNU	shavings_NN2 ._.
Long-Acre._NP1	Long-Acre_NP1 ._.
shopman._NNU	shopman_NN1 ._.
beer-shop._JJ	beer-shop_NN1 ._.
purport._NNU	purport_NN1 ._.
hair-dresser._JJ	hair-dresser_NN1 ._.
Deveil._NP1	Deveil_NP1 ._.
fleet-market._JJ	fleet-market_NN1 ._.
Alcock._NP1	Alcock_NP1 ._.
murder'd._NNU	murder'd_VVX ._.
police-office._JJ	police-office_NN1 ._.
frighted._NNU	frighted_JJ ._.
Egglestone._NP1	Egglestone_NP1 ._.
'Change._NNU	'Change_NN1 ._.
deficient._NNU	deficient_JJ ._.
wrapper._NNU	wrapper_NN1 ._.
concern'd._NNU	concern'd_JJ ._.
Panton._NP1	Panton_NP1 ._.
open'd._NNU	open'd_JJ ._.
tobacco-box._JJ	tobacco-box_NN1 ._.
Barron._NP1	Barron_NP1 ._.
Book-keeper._JJ	Book-keeper_NN1 ._.
Haydon._NP1	Haydon_NP1 ._.
George's-in-the-East._NP1	George's-in-the-East_NP1 ._.
truncheon._NNU	truncheon_NN1 ._.
inst._NNU	inst._NN1
Ratcliff._NP1	Ratcliff_NP1 ._.
Jacobs._NP1	Jacobs_NP1 ._.
stationer._NNU	stationer_NN1 ._.
Mile-end._JJ	Mile-end_NP1 ._.
Three-pence._JJ	Three-pence_NN1 ._.
errands._NNU	errands_NN2 ._.
beershop._NNU	beershop_NN1 ._.
ground-floor._JJ	ground-floor_NN1 ._.
before._NNU	before_CS. _.
counterpane._NNU	counterpane_NN1 ._.
waistcoat-pocket._JJ	waistcoat-pocket_NN1 ._.
afterward._NNU	afterward_RT ._.
Shew-glass._JJ	Shew-glass_NN1 ._.
damn'd._NNU	damn'd_VVX ._.
Fergusson._NP1	Fergusson_NP1 ._.
Mare._NP1	Mare_NN1 ._.
lost._NNU	lost_JJ ._.

Farrell._NP1	Farrell_NP1 ._.
search'd._NNU	search'd_VVX ._.
coal-cellar._JJ	coal-cellar_NN1 ._.
cheque-book._JJ	cheque-book_NN1 ._.
coat-pocket._JJ	coat-pocket_NN1 ._.
Tottenham-court-road._NP1	Tottenham-court-road_NP1 ._.
forecastle._NNU	forecastle_NN1 ._.
Gray's-inn-lane._NP1	Gray's-inn-lane_NP1 ._.
Homerton._NP1	Homerton_NP1 ._.
Koppel._NP1	Koppel_NP1 ._.
earrings._NNU	earrings_NN2 ._.
Deceas'd._NP1	Deceas'd_NN1 ._.
handkerchief._NNU	handkerchief_NN1 ._.
Selwyn._NP1	Selwyn_NP1 ._.
Watchman._NP1	Watchman_NN1 ._.
slaughter-house._JJ	slaughter-house_NN1 ._.
corps._NN	corps_NN1 ._.
Linnen._NP1	Linnen_NN1 ._.
pleases._NNU	pleases_VVZ ._.
Lodger._NP1	Lodger_NN1 ._.
interval._NN1	interval_NN1 ._.
Chick-Lane._NP1	Chick-Lane_NP1 ._.
approval._NN1	approval_NN1 ._.
trusses._NNU	trusses_NN2 ._.
crown-piece._JJ	crown-piece_NN1 ._.
dy'd._NNU	dy'd_VVD ._.
chopper._NNU	chopper_NN1 ._.
return'd._NNU	return'd_VVX ._.
over-board._NNU	over-board_RL ._.
watchmaker._NNU	watchmaker_NN1 ._.
own'd._NNU	own'd_VVX ._.
wife._NNU	wife_NN1 ._.
Work-house._JJ	Work-house_NN1 ._.
erased._NNU	erased_VVX ._.
says_VVZ I._NP1	says_VVZ I_PPIS1 ._.
as_CS33 I._NP1	as_CS33 I_PPIS1 ._.
as_II33 I._NP1	as_II33 I_PPIS1 ._.
than_CSN I._NP1	than_CSN I_PPIS1 ._.
do_VD0 I._NP1	do_VD0 I_PPIS1 ._.
did_VDD I._NP1	did_VDD I_PPIS1 ._.
N._NP1 B._NP1	N._ZZ1 B._ZZ1
Rev._NNU	Rev._NNB
th._NNU	th._MD
Augu_NN1 st._NNU	August_NPM1
cwt._FU	cwt._NNU

I_PPIS1 O_ZZ1 U._NP1	IOU_NN1 ._.
subp._NNU ena_NN1	subpoena_NN1
subp._NNU na_FU	subpoena_NN1
subp._NNU na_FW	subpoena_NN1
subp._NNU na._NNU	subpoena_NN1 ._.
subp._NNU na_UH	subpoena_NN1
subp._NNU naed_JJ	subpoenaed_JJ
subp._NNU naed._NNU	subpoenaed_JJ ._.
subp._NNU naed_VVD	subpoenaed_VVX
subp._NNU naed_VVN	subpoenaed_VVN
subp._NNU naing_JJ	subpoenaing_VVG
subp._NNU nas_NN2	subpoenas_NN2
subp._NNU nas._NNU	subpoenas_NN2
Ph._NN1 be_VBI	Phoebe_NP1
Ph._NN1 nix-court._NNU	Phoenix-court_NP1 ._.
Ph._NN1 nix_NN1	Phoenix_NP1
B--_NN1 h._NNU	B--h_NN1 ._.
b._NNU g._NNU y._NNU	b…g…y_NN1
b_ZZ1 -_- h._NNU	b--h_NN1 ._.
b._NNU h._NNU	b--h_NN1 ._.
h._NNU morrhage_NN1	haemorrhage_NN1
h._NNU morrhage._NNU	haemorrhage_NN1 ._.
h._NNU rrhage_NN1	haerrhage_NN1
h._NNU	h..._NN1
son_NN1 of_IO a_AT1 b._NNU	son_NN1 of_IO a_AT1 b..._NN1
b._NNU y_ZZ1	b...y_JJ
b._NNU ._. y_ZZ1	b…y_JJ
b._NNU r_ZZ1	b...r_NN1
b._NNU d_ZZ1	b...d_NN1
b._NNU h_ZZ1	b…h_NN1
b._NNU s_ZZ1	b…s_NN2
Q._NP1 C._NP1	Q.C._NNA
Q._NP1	Q._NN1