Dickens Dialect Project

Non-standard orthography, represented voice and characterisation across eleven works

Research presentation — expanded edition

11 WORKS

429 DISTINCT FORMS

729 CANDIDATE LINKS

A curated exploratory dataset. Not an exhaustive or token-verified inventory of Dickensian dialect.

© Tomoji Tabata (DH UOsaka)

Eleven works, one comparative framework

The project combines novels with a collection of short sketches; genres and speaking situations vary.

WORK	FIRST PUBLICATION	PG ID
Sketches by Boz	1833	#882
The Pickwick Papers	1836	#580
Oliver Twist	1837	#730
Nicholas Nickleby	1838	#967
Barnaby Rudge	1841	#917
Martin Chuzzlewit	1843	#968

WORK	FIRST PUBLICATION	PG ID
David Copperfield	1849	#766
Bleak House	1852	#1023
Hard Times	1854	#786
Little Dorrit	1855	#963
Our Mutual Friend	1864	#883

Years indicate the recorded first-publication start, not necessarily completion. Full bibliographical details should be verified for formal citation.

© Tomoji Tabata (DH UOsaka)

What has actually been counted?

Four distinct units of analysis must not be conflated.

11 **works**

Individual Project Gutenberg base texts

555 **work-specific form types**

A form is counted once within each work

429 **distinct orthographic forms**

Unique spellings after pooling the 11 works

729 **candidate form–speaker edges**

Proposed associations, not utterance tokens

302 / 729 edges carry a “sample-confirmed” label. This is evidence for a sampled form–speaker association, not verification of every occurrence.
The counts depend on the current curated inclusion list; absence from the list does not mean absence from Dickens.

© Tomoji Tabata (DH UOsaka)

“Dialect spelling” is not one linguistic variable

Separate phonology, eye dialect, vocabulary, morphosyntax, and deliberate character effects.

Phonetic respelling

wery ← very; thquire ← squire

Attempts to evoke perceived pronunciation.

Eye dialect

Non-standard spellings without a necessarily different pronunciation

Requires independent phonological justification.

Regional / social lexis

nowt; mun; local forms

Lexical identity may not be reducible to spelling.

Elision and contraction

reg’lar; p’raps; wi’

May be widespread in colloquial English.

Character-specific idiolect

demnition; umble

Stylistic signature rather than a geographic variety.

Performance / accommodation

“Me ope you leg well soon”

Represented speech adapted for a particular social interaction.

Labels are analytic hypotheses, not automatically recoverable facts about historical accents.

© Tomoji Tabata (DH UOsaka)

From source text to interpretable network

A sequential protocol with two mandatory human-review stages.

01 SOURCE

Fix edition, Project Gutenberg ID and provenance

02 CANDIDATE

Locate marked spellings; exclude ordinary forms

03 NORMALISE

Record original spelling + proposed standard equivalent

04 DISCOURSE

Identify quotation depth and spoken/written mode

05 SPEAKER

Assign individual speaker, confidence and passage

06 NETWORK

Aggregate validated edges; compare with denominators

The current site primarily supports stages 01–03 and candidate-level stage 05. Complete passage-level annotation is future work.

© Tomoji Tabata (DH UOsaka)

“Sample-confirmed” does not mean exhaustive

Keep exploratory evidence separate from a verified dialogue corpus.

DIMENSION	QUESTION TO ASK	EVIDENCE REQUIRED
Text occurrence	Does the form appear in the fixed source edition?	Full-text concordance check
Speaker evidence	Is at least one occurrence securely attributed?	Inspect reporting clauses + context
Quotation level	Is the form spoken here, quoted from a document, or nested?	Annotate discourse hierarchy
Token denominator	How many words does that speaker utter?	Extract and count verified direct-speech spans
Type frequency	How many form types are linked to a speaker?	Current graph: descriptive only

Recorded status: 302 sample-confirmed links; 427 links require speaker review. A named speaker–form edge should not be read as a frequency-weighted connection.

© Tomoji Tabata (DH UOsaka)

The Pickwick Papers

FIRST PUBLICATION: 1836–37

A conspicuous representation of London urban vernacular and comic oral storytelling.

Sam Weller

wery · vith · vich · p’raps

An abundance of /v/–/w/ spellings and oral idiom; reported anecdote and quoted speech can coexist.

Tony Weller

wos · vell · Samivel

Similar-looking features shared with Sam, yet differently distributed across family roles.

Analytic comparison

wery / vith / wot

Do not assume shared spelling entails identical pronunciation or social meaning.

“all I can say is, that I vish you may get it.” — Sam Weller

Pickwick candidate forms: 39; candidate links: 76. Whole-book spellings are not automatically Sam’s tokens.

© Tomoji Tabata (DH UOsaka)

Oliver Twist

FIRST PUBLICATION: 1837–39

Working-class and underworld voices are differentiated through markedly different orthographic devices.

Mr Gamfield

acause · chimbley · obstinit

Phonetic spellings in the chimney-sweep's portrayal; "gen'l'men" co-occurs in his speech.

Barney

bister · cub · dobody

Explicitly nasalised representation; should not be classified as a local accent without qualification.

Noah / Dodger

Work'us · wot · yer

School/workhouse and street idioms need speaker-level disambiguation.

"Bister Sikes! cub id, sir; cub id." — Barney

41 candidate forms; 64 proposed form–speaker links. Gender and class caricature are not direct dialect maps.

© Tomoji Tabata (DH UOsaka)

Nicholas Nickleby

FIRST PUBLICATION: 1838–39

The most elaborate Yorkshire-spelling cluster in the present candidate inventory.

John Browdie

thot · coom · loike · schoolmeasther

Forms associated with fictional Yorkshire speech; repeated across many scenes.

Yorkshire coach guard

hoold · whoam · wakken

Distinct speaker with overlapping phonetic patterns; must not be merged with Browdie.

Mr Mantalini

demmit · demnition · demnebly

A characteristic comic oath and idiolect, not Yorkshire dialect.

“Ye’ll be nowt but skeen and boans if you stop here long eneaf.” — John Browdie

183 candidate forms; 204 links. The Yorkshire informant belongs to Dickens’s Preface, not the fictional cast.

© Tomoji Tabata (DH UOsaka)

Martin Chuzzlewit

FIRST PUBLICATION: 1843–44

Personal mannerisms and socially situated vernacular complicate simple phonetic clustering.

Mrs Gamp

creetur · inwallable · sich

Gamp’s highly distinctive rhetoric frequently embeds quotations attributed to Mrs Harris.

Bailey

constitooshun · wot · wos

Printed pronunciation contributes to a youthful, irreverent social voice.

Mark Tapley

colloquial syntax + oral storytelling

Stylistic characterisation cannot be captured by non-standard spelling alone.

“you are that inwallable person.” — quoted Mrs Harris, as narrated by Mrs Gamp

29 candidate forms; 47 links. The inwallable example is nested quoted discourse, a deliberate attribution trap.

© Tomoji Tabata (DH UOsaka)

David Copperfield

FIRST PUBLICATION: 1849–50

Distinct dialect stylisation, self-presentation and social positioning within one first-person novel.

Uriah Heep

umble · umblest · umbleness

Recurrent h-dropping-like self-description performs obsequiousness and calculated humility.

Mr Peggotty

wot · obleeged · Dan’l

Nautical/rural speech and affectionate kinship naming in Norfolk-related scenes.

Mrs Gummidge / Barkis

creetur / willin’ · writin’

Formulaic self-description and habitual expressions serve different character functions.

“I’m a very umble person.” — Uriah Heep

36 candidate forms; 50 links. Beware narratorial repetition of a character’s catchphrase.

© Tomoji Tabata (DH UOsaka)

Bleak House

FIRST PUBLICATION: 1852–53

Representation of poverty, institutional power and restricted linguistic access.

Jo

nothink · wos · wery

Emblematic street-child voice; patterns also appear in narratorial framing, not only direct dialogue.

Phil Squod

guv’ner · colloquial reductions

A distinct veteran’s voice: phonetic effects overlap with broader urban non-standard forms.

Coavinses

know’d · yes’day · aft’noon

Bailiff-related speech with a characteristic working-class voice.

“For I don’t … I don’t know nothink.” — Jo

45 candidate forms; 61 links. Narrated descriptions of Jo’s understanding are not direct speech.

© Tomoji Tabata (DH UOsaka)

Our Mutual Friend

FIRST PUBLICATION: 1864–65

Compare riverine and marginalised urban characters without treating them as one homogeneous Cockney type.

Rogue Riderhood

reg’lar · arter · p’raps

Speech combines non-standard form, rhetorical defensiveness and strategic self-presentation.

Sloppy

wos · wery

A different character pattern; one form may surface within someone else’s account of him.

Silas Wegg

performance and inflated recitation

Reading aloud and deliberately florid diction require an additional performance category.

“I never heerd it called that … p’raps you call it so.” — Riderhood

33 candidate forms; 45 links. Speaker review is incomplete for most form–speaker pairs.

© Tomoji Tabata (DH UOsaka)

Barnaby Rudge

FIRST PUBLICATION: 1841

Dramatic political-historical fiction: voice can perform occupation, power and class simultaneously.

Ned Dennis

wery · helegant · dewoted

Hypercorrect-sounding or non-standard forms help stage the hangman's comic self-image.

Miss Miggs

socially marked complaint / affect

Spellings alone cannot represent the full register of grievance and interpersonal performance.

John Willet / Sim Tappertit

selected provincial or comic forms

Candidate associations should be verified against surrounding narrative voice.

"the helegant bits of work it has turned off" — Ned Dennis

19 candidate forms; 27 links. Avoid inferring modern phonemic values directly from spelling.

© Tomoji Tabata (DH UOsaka)

Hard Times

FIRST PUBLICATION: 1854

A central methodological test: idiosyncratic articulation is not synonymous with northern regional dialect.

Mr Sleary — lisp

thquire · thervant · thith · piethe

The th/s substitutions are a characterised speech impediment, not a regional accent label.

Stephen Blackpool

mun · coom · wi’ · awlus

Northern working-class dialect stylisation operates through different morphophonological cues.

Rachael

shared local forms, distinct voice

A common non-standard form should not erase variation among individual workers.

“Thquire! Your thervant! Thith ith a bad piethe of bithnith.” — Sleary

64 candidate forms; 74 links. The two-character contrast warrants independent annotation categories.

© Tomoji Tabata (DH UOsaka)

Little Dorrit

FIRST PUBLICATION: 1855–57

Sparse marked spellings span prison-related settings, service work and accommodative speaking.

Mr Plornish

wery · ed

Urban artisan speech with possible h-loss or non-standard pronouncing conventions.

Marshalsea turnkey

p’raps · ackney

Contractions and dropped initial h in reported working-world exchanges.

Mrs Plornish / Maggy

“Me ope you leg well soon” / reg’lar

Foreigner-directed simplification and more ordinary colloquialism must be separated.

“Me ope you leg well soon.” — Mrs Plornish, in an accommodated speech context

24 candidate forms; 29 links. Identical-looking forms have different discourse functions across speakers.

© Tomoji Tabata (DH UOsaka)

Sketches by Boz

FIRST PUBLICATION: 1833–36

The sampling unit is often the sketch or episode, not a recurring novel character.

Omnibus cad / coachmen

werry · vith · vich

Occupationally embedded voices and brief encounters in metropolitan transport.

Crofts / Samuel Wilkins

sich · wery / wos

Named individual voices occur, but the whole collection is not a fixed social network.

Anonymous street voices

niver · colloquial reductions

Preserve a local event ID or sketch heading rather than making a fictional pooled “speaker”.

“I wos a thinking of taking J’mima …” — Samuel Wilkins

42 candidate forms; 52 links. Multi-sketch aggregation can conceal contextual variation.

© Tomoji Tabata (DH UOsaka)

What different forms are doing

A working classification of a *function*, not a claim that each character speaks one dialect.

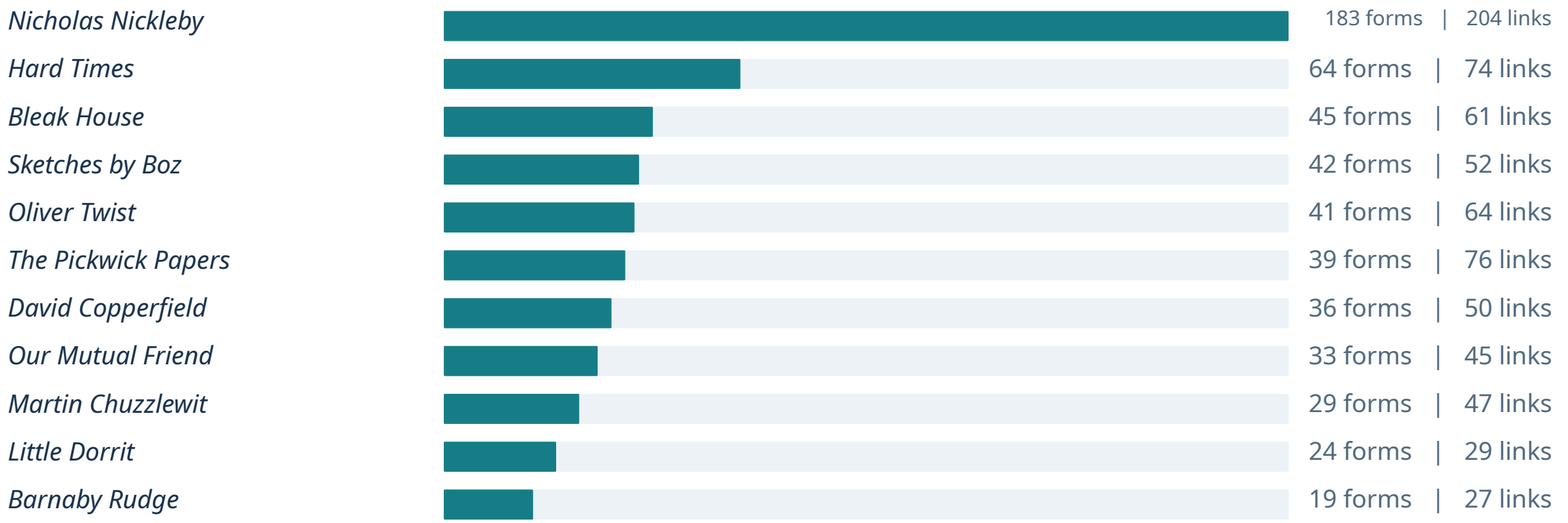
INTERPRETIVE MODE	SPEAKERS	FORMS	FUNCTION / CAUTION
Regionalised voice	Browdie / Stephen Blackpool	thot / coom / mun	Candidate regional phonological and lexical cues
Urban colloquial	Sam Weller / Jo / Riderhood	wery / nothink / p’raps	Overlapping urban, social and represented voice
Idiolect / oath	Mantalini / Uriah Heep	demnition / umble	Distinctive habitual characterisation
Articulation stylisation	Sleary / Barney	thquire / bister	Lisp versus nasal speech portrayal
Accommodation / performance	Mrs Plornish / Wegg	Me ope / ornate recitation	Not equivalent to speaker’s ordinary dialect

Some forms also occur in narratorial prose, embedded texts or the speech of multiple characters.

© Tomoji Tabata (DH UOsaka)

Candidate inventory, by work

Distinct form types (bars) and associated form–speaker links (labels). Values reflect curation, not dialect density.



Do not compare “dialect frequency” across books without utterance token counts, normalisation and harmonised inclusion criteria.

© Tomoji Tabata (DH UOsaka)

A quotation does not identify a speaker

Nested reporting can make the same typography misleading for automated attribution.

LEVEL 1

Dickens’s narrator or first-person memoir

Storytelling framework: dialogue may be narrated, represented or summarised.

LEVEL 2

One character’s direct speech

Mrs Gamp speaks; she may recount a past conversation.

LEVEL 3

Quoted speech inside that account

“inwallable” is attributed to Mrs Harris in Gamp’s anecdote.

Narrator → character’s speech → character’s recollected quotation

Record the current utterer and the represented earlier utterer separately; never assign every quoted token to the outermost speaker.

© Tomoji Tabata (DH UOsaka)

Keep form, mode, speaker and confidence independent

Example of a hierarchical research encoding (illustrative XML; not an existing annotation).

```
<utterance who="#Gamp" mode="spoken"
          embedding-depth="1">
  … <quote who="#MrsHarris"
            mode="recalled-spoken"
            embedding-depth="2">
       <w orig="inwallable" norm="invaluable"/>
     </quote> …
</utterance>
```

Token	Original orthography; case-sensitive display + normalisation
Context	Work, chapter/episode, passage offsets, edition
Discourse	Spoken / written / thought / quoted / ambiguous
Relation	Current speaker; source of embedded quotation
Review	Independent attribution, confidence, review notes

A printed quotation mark alone is not evidence of direct speech: speech, thought and written documents must be separated.

© Tomoji Tabata (DH UOsaka)

Validation and reproducibility checklist

What needs to happen before making quantitative claims about Dickensian dialect?

01 FIX TEXTS

Record edition, file hash, encoding and version.

02 EXHAUSTIVE SEARCH

Retrieve candidates by concordance + orthographic variants.

03 INDEPENDENT REVIEW

Double-annotate uncertain speaker and discourse cases.

04 NORMALISE COUNTS

Per-speaker utterance tokens; rates per 1,000 words.

05 ROBUSTNESS

Sensitivity to category definitions and weak evidence.

06 RELEASE

Versioned CSV/JSON, documentation and transparent exclusions.

Research questions: Which characters share forms? Which spellings are highly distinctive after exposure correction? Where is “accent” actually an idiolect or performed voice?

© Tomoji Tabata (DH UOsaka)

Source texts, provenance and next step

The current public-facing graph is an exploratory index into the literary evidence.

SOURCE TEXTS / PROJECT GUTENBERG

The Pickwick Papers	#580
Oliver Twist	#730
Nicholas Nickleby	#967
Barnaby Rudge	#917
Martin Chuzzlewit	#968
David Copperfield	#766
Bleak House	#1023
Hard Times	#786
Little Dorrit	#963
Our Mutual Friend	#883
Sketches by Boz	#882

INTERPRETATION NOTES

1. All counts derive from the project's bundled dataset (11-work edition).

2. The 302 "sample-confirmed" labels do not establish token-complete annotation.

3. Read examples in their original textual context, especially reported and embedded speech.

Public HTML companion: slides.html · Methodology: methodology.html · Network: index.html

© Tomoji Tabata (DH UOsaka)