Encyclopaedia of Shakespeare's Language

Lancaster University

# Grappling with Shakespeare's words: maximizing historical corpus-based approaches

**Jonathan Culpeper and Amelia Joulain-Jay**

**Lancaster University**

@ShakespeareLang

http://wp.lancs.ac.uk/shakespearelang/

Arts & Humanities Research Council

CASS Corpus Approaches to Social Science

The Queen's Anniversary Prizes For Higher and Further Education 2015

UCREL

# The Encyclopedia of Shakespeare's Language project: A corpus-based approach to Shakespeare

Lancaster University

- For Shakespeare's language this means:
  - All 'words' treated equally (not just 'hard' words).
  - Meanings based on usage in context (not etymology, not narrowly-defined semantic meaning).
  - Context includes linguistic aspects (e.g. collocations) and non-linguistic aspects (e.g. registers, social properties of speaker/character).
  - A comparative approach.

# What will be in the encyclopaedia?

Lancaster University

- Volume 1 (a kind of dictionary)
  - The use and meanings of each of Shakespeare's words, both in the context of what he wrote and in the context in which he wrote.
  - Every word is, for example, compared with a 321 million word corpus comprising the work of Shakespeare's contemporaries.
  - Establishes both what is unique about Shakespeare's language and what Shakespeare's language meant to his contemporaries.

# What will be in the encyclopaedia?

- Volume 2 (a kind of compendium of semantic patterns)

➢ **Plays and characters**
  - Major characters (> 5% of total word count)
  - Play profiles
  - Comedies
  - Histories
  - Tragedies

➢ **Gender and social stratification**
  - Male/female
  - Social status

➢ **Major themes in Shakespeare**
  - Love and marriage
  - War and conflict
  - etc.

➢ **Genre**

# Our core data: Shakespeare texts

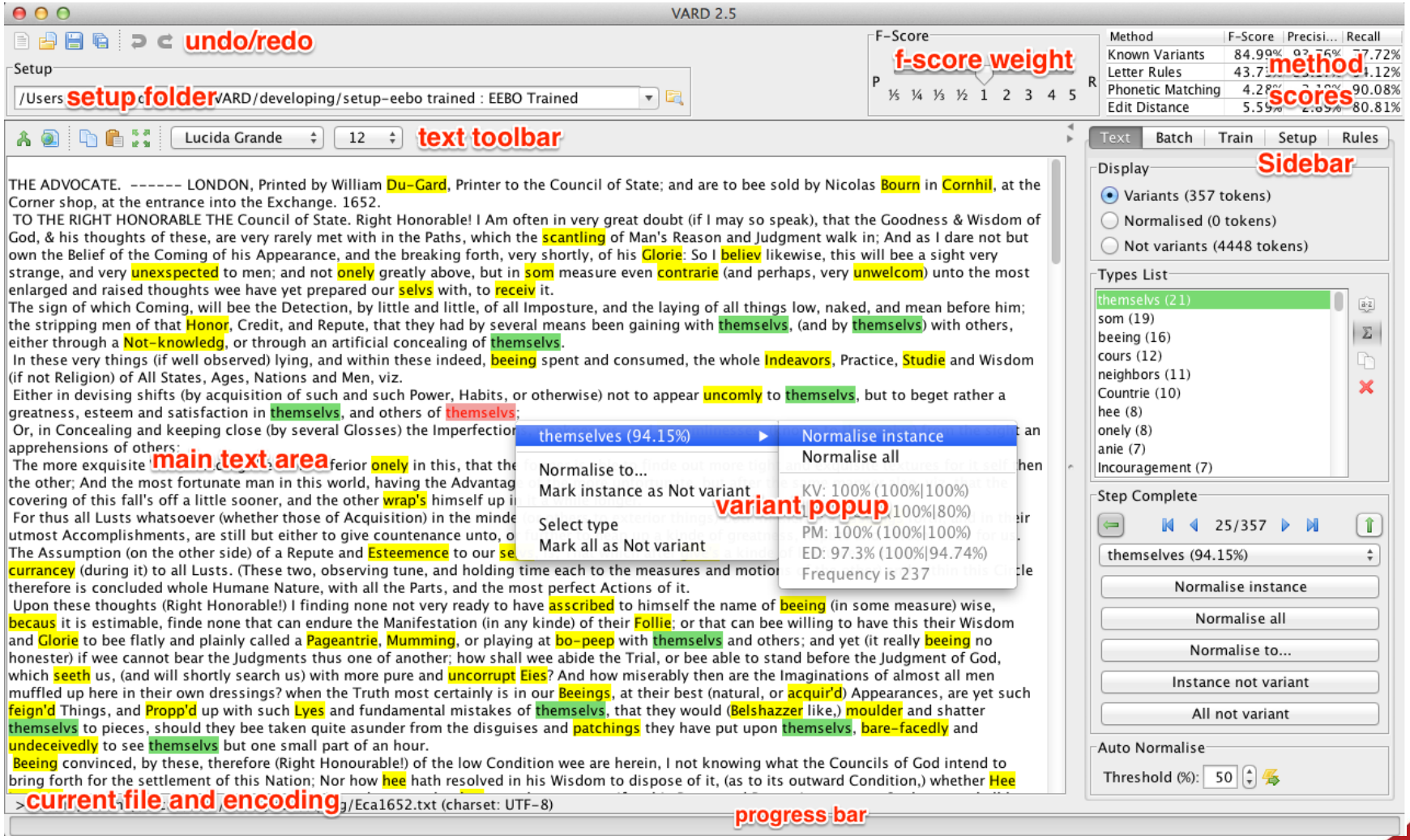
- Core data: plays generally agreed to be part of the Shakespeare canon
  - The largest near-contemporary body of work attributed to Shakespeare, i.e. the First Folio (1623), plus *Pericles* and *The Two Noble Kinsmen*.
- Quartos constitute a secondary dataset.
- Poetry constitutes a third Shakespeare dataset.

# Spelling variation

- If untreated, frequency counts for a word would be split across several variants
  - e.g. in our core Shakespeare data, *would* is also spelled *vvould*, *wold*, *wad*
- Addressing this improves the prospects for matching words in target and reference corpora
- Easier for the present-day reader/user
- Our solution: use VARiant Detector (VARD 2) software (Baron & Rayson 2008; http://ucrel.lancs.ac.uk/vard/)

# VARD 2 (VARiant Detector)

http://ucrel.lancs.ac.uk/vard/



# Spelling variation

Lancaster University

**Problem**: What do you regularize the spelling to? There is no standardised regular form in the way that there is today.

**Solution**: Our policy was to

- Preserve the morphology, e.g. 2nd + 3rd person verb inflections (–(e)st, -(e)th), past tense forms (e.g. *holp*), past participle forms (e.g. *holpen*), plural forms (e.g. *shooen*), non-standard superlatives (e.g. *horrider*), and *you*/*thou*,
- Only use a spelling that had EModE currency.
- Prioritize the most frequent spelling in Shakespeare

# Part of speech tagging and EModE (cf. Andrew Hardie)

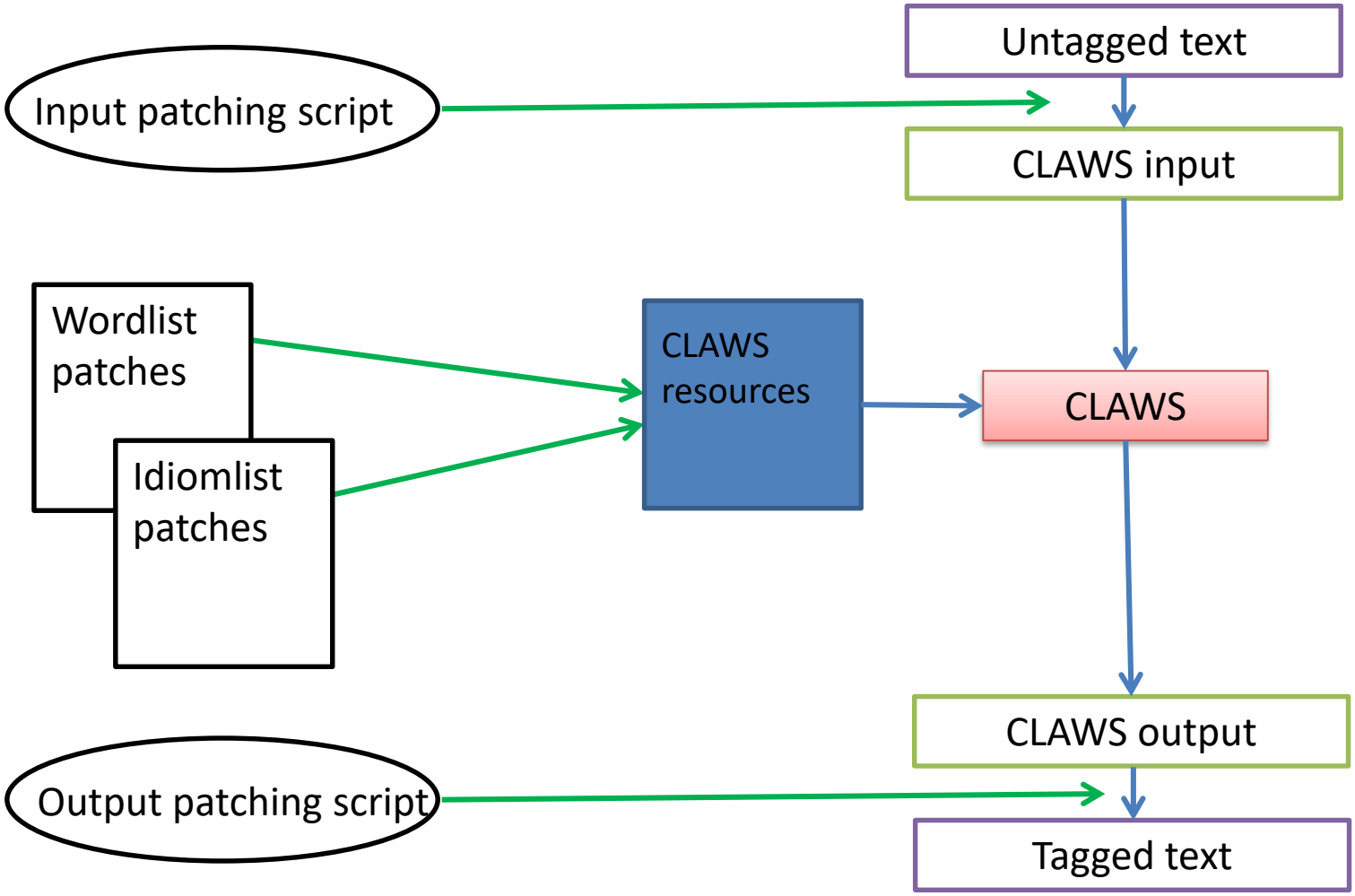
Lancaster University

CLAWS performs at 85% accuracy for Shakespearean texts (Rayson et al. 2007). Not good enough!

- Grammatical phenomena which are marginal today, but may require addressing by the tagger for EModE (e.g. 2nd person singular)
- Words not in the modern lexicon (e.g. *hent*)
- Words whose possible classifications have changed (e.g. *faith*)
- Words whose probability profile has changed (e.g. *prostitute*)
- Extra cliticisations (e.g. me=thinks, me=thought)

- No time for a radical system overhaul
- Solution: Patches for CLAWS

# Implementation



# Data and genre: *Early English Books Online (TCP)* 1560-1639 (379 million words; 5,750 texts categorized by genre, domain and style)

Lancaster University

| Styles | Domains | Genres | Sub-genres (examples) |
|---|---|---|---|
| Literary | Imaginative | Plays<br>Poetry, Verse & Song<br>Fiction<br>General | Comedy, History, Tragedy, Masque<br>Ballads |
| Formal – Spiritual | Religion | Bible<br>Catholicism<br>Protestantism<br>Doctrine, Theology and Governance<br>General | <br>Anti-Catholicism<br>Church of England<br>Sin and Repentance<br>Sermons |
| Formal - Statutory | Government | Royal<br>Parliamentary<br>Legal<br>General | Proceedings<br>Reports<br>Trials<br>Speeches |
| Formal - Instructional | Didactic | Philosophy<br>Science<br>Mathematics<br>Medicine<br>General | <br>Experiments<br><br>Anatomy<br>Alchemy |
| Informational | Factual | Biography<br>Essay<br>Letters<br>Pamphlets<br>General | Dialogue<br><br><br><br>Food and Cookery |

# The challenge

Multiple pieces of information:

- spellings, parts-of-speech, collocates, genre distribution, social distribution (e.g. male/female; high rank/low rank)

In multiple information sets:

- Shakespeare's plays, his poetry, the Folios, the Quartos, our comparative corpus of playwrights and the EEBO-TCP.

# Lexicography interface: background

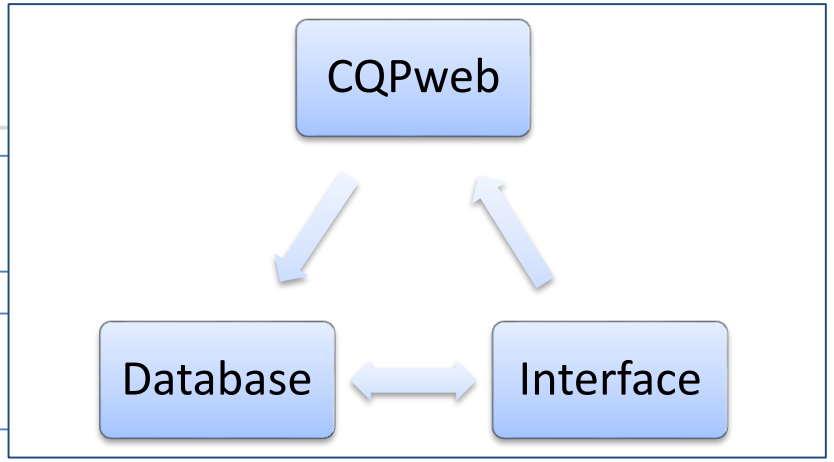
Lancaster University

- We have CPQweb… why have another interface?
  - automation of repetitive tasks
  - dictionary writing system

- What does the system involve?
  - a MySQL database with two parts:
    - fixed data (number of occurrences, etc.)
    - modifiable data (definitions, etc.)
  - a user interface which:
    - provides access to the data
    - allows for the creation of modifiable data

# Lexicography interface: set-up

Lancaster University




- Corpus is pre-processed
- Corpus is uploaded to CQPweb
- Data is extracted from CQPweb and stored in the fixed part of the database
- Interface displays fixed data
- Users interact with the interface to create data which gets stored in the modifiable part of the database
- Interface displays fixed and modifiable data

# Lexicography interface: demo

https://corpora.lancs.ac.uk/shakencyc/

# A more grammatical word: *I*

**Shakespearean dictionaries**:

- Words such as this omitted from Shakespearean dictionaries (e.g. Crystal and Crystal 2002; Onions 1986), presumably on the assumption that they:

  (a) have obvious meanings (because they are considered more or less the same as those of today), and

  (b) do not contribute much to understanding Shakespeare.

# A more grammatical word: *I*

Lancaster University

## Encyclopaedia of Shakespeare's Language

### Definition preview: definition 122 for entry i_PRON

**i******* *pron.* (I, me):
As now. Most regularly co-occurs with verbs, relating to mental states/processes (e.g. HOPE, THINK, KNOW, REMEMBER, FORGET, DREAM, UNDERSTAND, DOUBT, WISH) and speech act devices (e.g. PRAY, THANK, SWEAR, BESEECH, ASSURE, PROMISE, BEG, ENTREAT, MEAN). Co-occurs frequently with verbs relating to communication (e.g. TELL, SAY, READ, WRITE), movement (e.g. GIVE, TAKE, GO), emotion state/processes (e.g. FEAR, LOVE, HATE) and perception (e.g. SEE, HEAR, WATCH). Aside from modal verbs (e.g. CAN, WILL, MAY) and adverbs, especially involving negation (e.g. NEVER, NOT, NO), it co-occurs regularly with the pronouns THOU, YOU, MY and IT.

T C H

# *I* and Desdemona

Lancaster University

## Desdemona's keywords

| | Raw freq. | Log-L. | LogRatio |
|---|---|---|---|
| prithee | 8 | 16.47 | 3.24 |
| lord | 39 | 64.82 | 2.74 |
| lost | 7 | 10.4 | 2.53 |
| alas | 8 | 8.7 | 2.04 |
| him | 41 | 24.75 | 1.41 |
| do | 44 | 19.64 | 1.18 |
| my | 79 | 28.03 | 1.03 |
| me | 47 | 11.61 | 0.84 |
| i | 132 | 26.85 | 0.76 |

For Othello: *I* is ranked 109, *me* 70 and *my* 74

# A more lexical word: *good*

Lancaster University

Dictionaries (in brief):

- *Onions* (1911): (1) Conventional epithet to titles of high rank, (2) comely, (3) Financially sound; (hence) wealthy, substantial.

- *Crystal & Crystal* (2004): (1) [intensifying use] real, genuine ('love no man in good earnest'). (2) kind, benevolent, generous. (3) kind, friendly, sympathetic. (4) amenable, tractable, manageable. (5) honest, virtuous, honourable. (6) seasonable, appropriate proper. (7) just, right, commendable. (8) intended, right, proper. (9) high-ranking, highborn, distinguished. (10) rich, wealthy, substantial.

# A more lexical word: *good*

Lancaster University

## Encyclopaedia of Shakespeare's Language

### Definition preview: definition 104 for entry good_ADJ

**good****** *adj.* (good, better, best):
1. A polite address: '(my) good Lord/friend/Sir/Master/Lady/Madam/etc.'. Typically used when meeting or parting, thanking or making suggestions. *But (good my Lord) do it so cunningly* TGV, III. 1.
2. Honest, truthful, principled; of high moral standards. (This sense also shapes the discourse markers '(in) good faith/sooth/troth', which mean truly or honestly). *a man of good repute, carriage, bearing, & estimation* LLL, I. 1.
3. Positive rather than negative. Typically, contrasted with 'bad'. *Is thy news good or bad?* ROM, II. 5.
4. In one's favour, especially favourable wishes or blessings. *The Gods be good to us* COR, V. 4.
5. A welcoming, cheerful manner. *Therefore for Gods sake entertain good comfort, And cheer his Grace with quick and merry eyes* R3, I. 3.
**good will** As now.
**good morrow** Good morning.
**good night** As now.

T C H

Lancaster University

# Questions?

Lancaster University

# Input patching

Lancaster University

```
10 // splits that CLAWS does for us already:
11 // twas
12 // tis
13 // twould
14 // twere
15
16 $splits_raw = <<<END
17 me|thinks
18 me|thought
19 me|thoughts
20 me|seems
21 me|seemeth
22 me|seemed
23 t|as
24 t|will
25 END;
```

# EModE tags

| Tag | Words |
|---|---|
| PPYS1 | thou, th' |
| PPYO1 | thee |
| VBT | art, beest |
| VBDT | wast, wert |
| VDT | dost, doest |
| VDDT | didst |
| VHT | hast |
| VHDT | hadst |
| VMT | wilt, wouldst, canst, couldst, shalt, shouldst etc. |
| VMTK | oughtest, usedest |
| VVT | givest, workest |
| VVDT | gavest, workedst |

# Resource patching: lexicon

```
28 nill VM
29 wilt VM VV0@ NN1%
30 wouldst VM
31 canst VM
32 couldst VM
```

```
11 art VBR NN1@ NP1:%
12 beeth VBZ
13 beest VBR
14 wast VBDR
15 wert VBDR
```

```
76 calledest VVD
77 calledst VVD
78 callest VV0
79 calleth VVZ
80 callst VV0
81 camest VVD
82 camst VVD
83 comest VV0
84 cometh VVZ
85 comst VV0
86 cutteth VVZ
```

```
404 / some extra nonlex adverbs
405 thereat RR
406 hereat RR
407 therewithal RR
408 herewithal RR
409 thitherward RL
410 hitherward RL
411 thitherwards RL
412 hitherwards RL
```

```
358 unworthier JJR
359 verier JJR
360 violentest JJT
361 welcomest JJT
362 wholesomest JJT
363 willingest JJT
364 wiselier JJR
365 woefullest JJT
366 worser JJR
```

```
289 bid NN1 VV0 VVN VVD
290 bids NN2 VVZ
291 fee NN1 VV0
292 fees NN2 VVZ
293 feed VV0 NN1 VVD@ VVN@
294 prostitutes VVZ NN2@
295 prostitute VV0 NN1@
296 passing VVG JJ RR NN1@
297 marry VV0 UH
298 wanton VV0 JJ NN1
299 wantons VVZ NN2
```

```
420 for IF CS
421 but CCB RR II
422 an AT1 CS%
```

```
436 /intejections, disc particles
437 o UH ZZ1@
438 fie UH
439 faith NN1 UH
440 ifaith UH
441 i'faith UH
442 inprimis RR
443 iwis RR
444 prithee UH VV0@
```

# Output patching

```
/ pronouns
thou PPY PPYS1
thee PPY PPYO1
th PPY PPYS1
th' PPY PPYS1

/ verbs
art VBR VBT
beest VBR VBT
wast VBDR VBDT
wert VBDR VBDT
hast VH0 VHT
havest VH0 VHT
hadst VHD VHDT
haddest VHD VHDT
dost VD0 VDT
doest VD0 VDT
didst VDD VDDT
wilt VM VMT
wouldst VM VMT
canst VM VMT
```

```
seemedst VVD VVDT
seemest VV0 VVT
seemst VV0 VVT
seest VV0 VVT
showedest VVD VVDT
showedst VVD VVDT
showdst VVD VVDT
showest VV0 VVT
showst VV0 VVT
showst VV0 VVT
startedest VVD VVDT
startedst VVD VVDT
startest VV0 VVT
startst VV0 VVT
takest VV0 VVT
takst VV0 VVT
talkedest VVD VVDT
talkedst VVD VVDT
```

```
owst VV0 VVT
scaldst VV0 VVT
scoldst VV0 VVT
scornst VV0 VVT
scorndst VVD VVDT
seekst VV0 VVT
sentst VVD VVDT
servst VV0 VVT
setst VV0 VVT VVD VVDT
settlest VV0 VVT
shakst VV0 VVT
shamst VV0 VVT
shinst VV0 VVT
shrugst VV0 VVT
```