Comprehensive Annotation of Multiword Expressions in a Social Web Corpus

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik,
Michael T. Mordowanec, Henrietta Conrad, Noah A. Smith

LREC • May 28, 2014

CMWE, a text corpus comprehensively annotated with **multiword expressions**

CMWE, a text corpus comprehensively annotated with **multiword expressions**

- ✦ 55k words of English web reviews

CMWE, a text corpus comprehensively annotated with **multiword expressions**

- ✦ 55k words of English web reviews
- ✴ fully heterogeneous MWEs

CMWE, a text corpus comprehensively annotated with **multiword expressions**

- ✦ 55k words of English web reviews
- ✴ fully heterogeneous MWEs
- ✴ shallow groupings, allowing gaps

CMWE, a text corpus comprehensively annotated with **multiword expressions**

- ✦ 55k words of English web reviews
- ✴ fully heterogeneous MWEs
- ✴ shallow groupings, allowing gaps
- ✴ strong vs. weak

CMWE, a text corpus comprehensively annotated with **multiword expressions**

- ✦ 55k words of English web reviews
- ✴ fully heterogeneous MWEs
- ✴ shallow groupings, allowing gaps
- ✴ strong vs. weak

tinyurl.com/mwecorpus

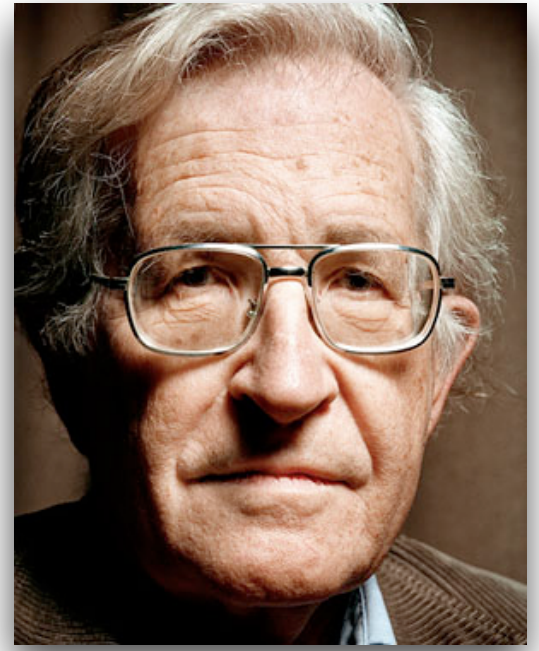
Outline

Outline

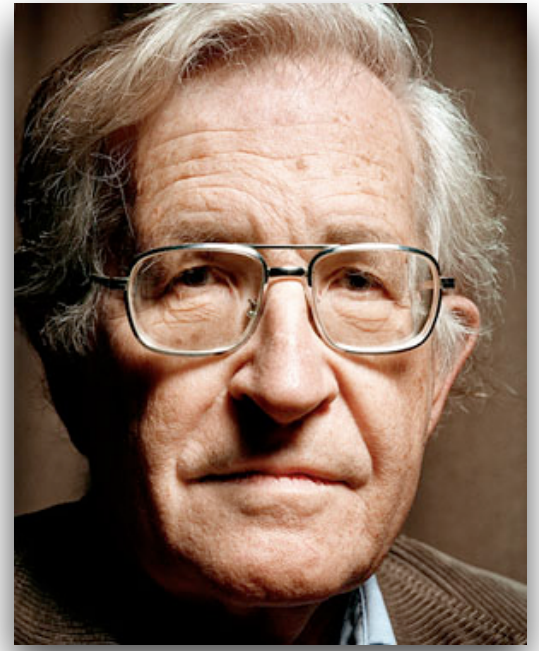
- multiword expressions
- annotation process
- the corpus

daddy longlegs

daddy longlegs



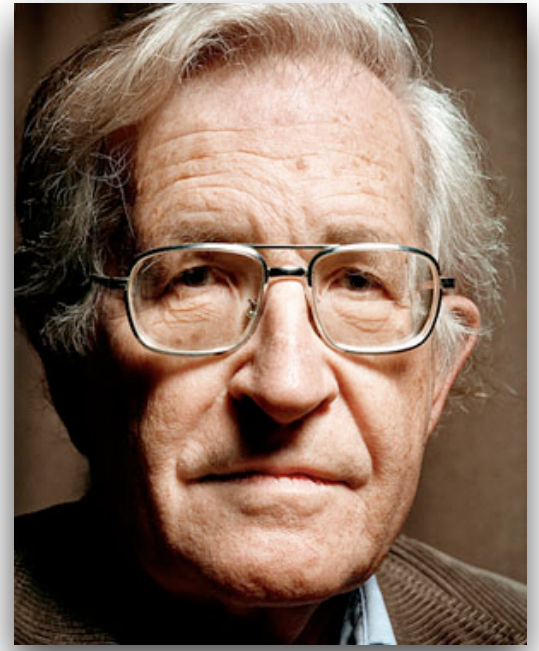
Noam Chomsky →



daddy longlegs

↘

Noam Chomsky →



daddy longlegs

↘

The aliens will destroy Earth unless we

their demands.

The aliens will destroy Earth unless we

accept

their demands.

The aliens will destroy Earth unless we

accept

agree to

accede to

yield to

give in to

their demands.

The aliens will destroy Earth unless we

accept

agree to
accede to
yield to
give in to

comply with
cooperate with
go along with

their demands.


Jonathan Huang

Jonathan Huang

Noam Chomsky refused to give in to the vicious daddy longlegs .

Noam_Chomsky refused to give in to the vicious daddy longlegs .

Noam_Chomsky refused to give_in_to the vicious daddy longlegs .

Noam_Chomsky refused to give_in_to the vicious daddy_longlegs .

Noam_Chomsky
refused
to
give_in_to
the
vicious
daddy_longlegs
.

Lexical segmentation

Noam_Chomsky

refused

to

give_in_to

the

vicious

daddy_longlegs

.

Definition

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency
  - *unusual morphosyntax*: **Me/*Him neither**; **by and large**; plural of **daddy longlegs**?

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency
  - *unusual morphosyntax*: **Me/*Him neither**; **by and large**; plural of **daddy longlegs**?
  - *non-* or *semi-compositional*: **ice cream**, **daddy longlegs**, **pay attention**

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency
  - *unusual morphosyntax*: **Me/*Him neither**; **by and large**; plural of **daddy longlegs**?
  - *non-* or *semi-compositional*: **ice cream**, **daddy longlegs**, **pay attention**
  - statistically *collocated*: *p*(**highly unlikely**) > *p*(**strongly unlikely**)

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency
  - *unusual morphosyntax*: **Me/*Him neither**; **by and large**; plural of **daddy longlegs**?
  - *non-* or *semi-compositional*: **ice cream**, **daddy longlegs**, **pay attention**
  - statistically *collocated*: *p*(**highly unlikely**) > *p*(**strongly unlikely**)

Definition

- **Multiword expression** (MWE): 2 or more orthographic words/lexemes that function together as an **idiomatic whole**
- *idiomatic* = not fully predictable in form, function, and/or frequency
  - *unusual morphosyntax*: **Me/*Him neither**; **by and large**; plural of **daddy longlegs**?
  - *non-* or *semi-compositional*: **ice cream**, **daddy longlegs**, **pay attention**
  - statistically *collocated*: *p*(**highly unlikely**) > *p*(**strongly unlikely**)

Challenges

Challenges

- Not superficially apparent in text
- Number/frequency
  - Too many expressions to list all of them
  - Individually rare, but frequent in aggregate
- Diversity
  - Many different construction types
  - Semantically unrestricted
  - Can be **gappy**

Fred Jelinek

daddy longlegs, hot dog

dry out

depend on, come across

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where __ left off

easy as pie

You're welcome.

To each his own.

The structure of this paper is as follows.

Fred Jelinek

daddy longlegs, hot dog

dry out

depend on, come across

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where __ left off

easy as pie

You're welcome.

To each his own.

structure of this paper is as follows.

dry out

pay attention (to)

pick up where __ left off

Fred Jelinek

daddy longlegs, hot dog

dry out

depend on, come across

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where __ left off

easy as pie

You're welcome.

To each his own.

structure of this paper is as follows.

dry out the clothes

pay attention (to)

pick up where __ left off

Fred Jelinek

daddy longlegs, hot dog

dry out

depend on, come across

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where __ left off

easy as pie

You're welcome.

To each his own.

structure of this paper is as follows.

dry the clothes **out**

pay attention (to)

pick up where __ left off

Fred Jelinek

daddy longlegs, hot dog

dry out

depend on, come across

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where __ left off

easy as pie

You're welcome.

To each his own.

structure of this paper is as follows.

dry the clothes **out**

pay close **attention (to)**

pick up where __ left off

Fred Jelinek

daddy longlegs, hot dog

dry out

depend on, come across

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where __ left off

easy as pie

You're welcome.

To each his own.

structure of this paper is as follows.

dry the clothes **out**

pay close **attention (to)**

no **attention** was **paid (to)**

pick up where __ left off

Fred Jelinek

daddy longlegs, hot dog

dry out

depend on, come across

pay attention (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where __ left off

easy as pie

You're welcome.

To each his own.

structure of this paper is as follows.

dry the clothes **out**

pay close **attention (to)**

no **attention** was **paid (to)**

pick up where they **left off**

Examples

My wife had taken_ her '07_Ford_Fusion _in for a routine oil_change .

Examples

They gave me the run around and missing paperwork only to call back to tell me someone else wanted her and I would need to come in and put down a deposit .

Examples

They gave_ me _the_run_around and missing paperwork only to call_back to tell me someone else wanted her and I would need to come_in and put_down~ a ~deposit .

Examples

They gave_ me _the_run_around and missing paperwork only to call_back to tell me someone else wanted her and I would need to come_in and put_down~ a ~deposit .

weak MWE (collocation); cf. highly~recommended

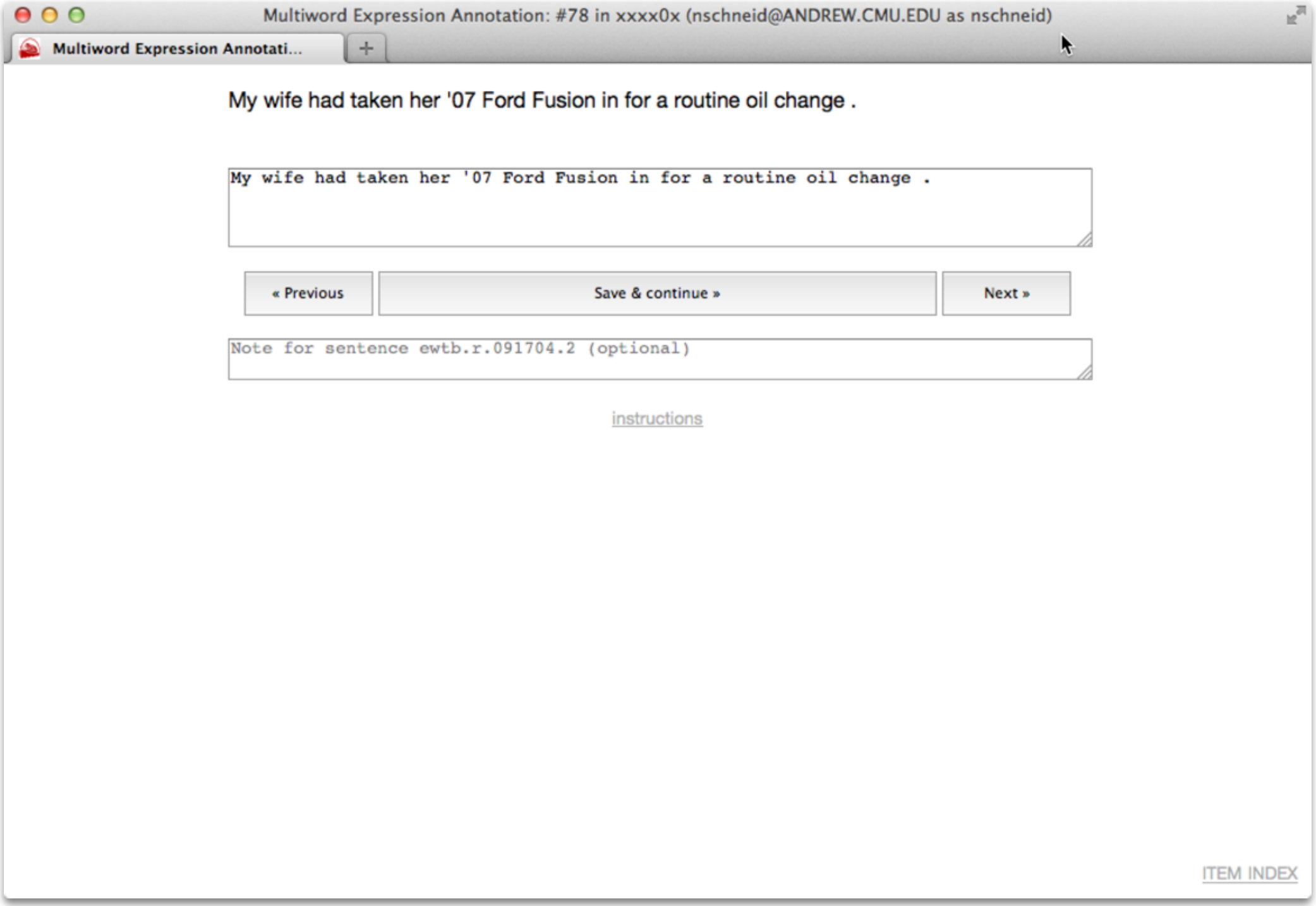
Examples

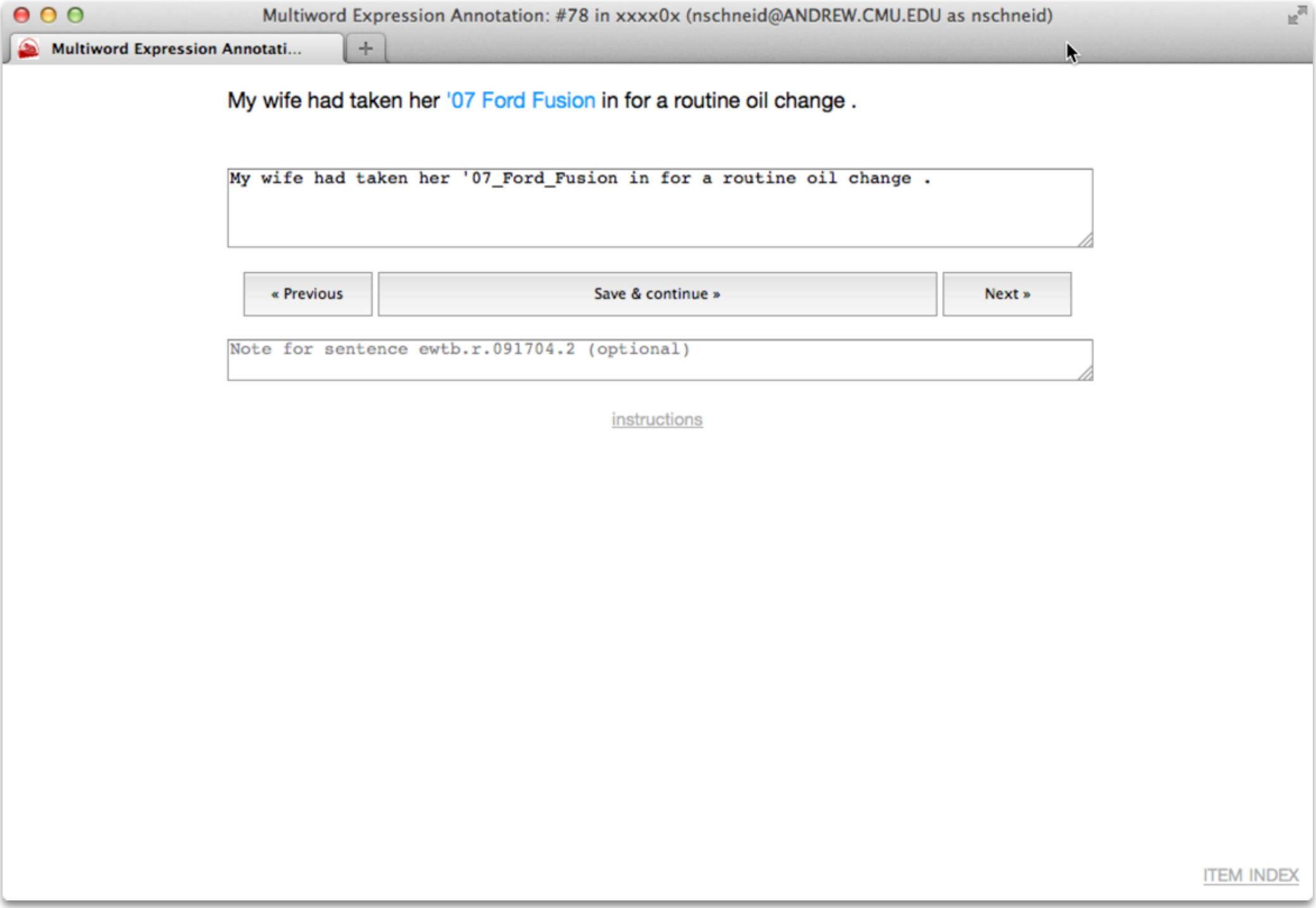
Among the animals that were available to touch were pony's , camels and EVEN AN OSTRICH !!!

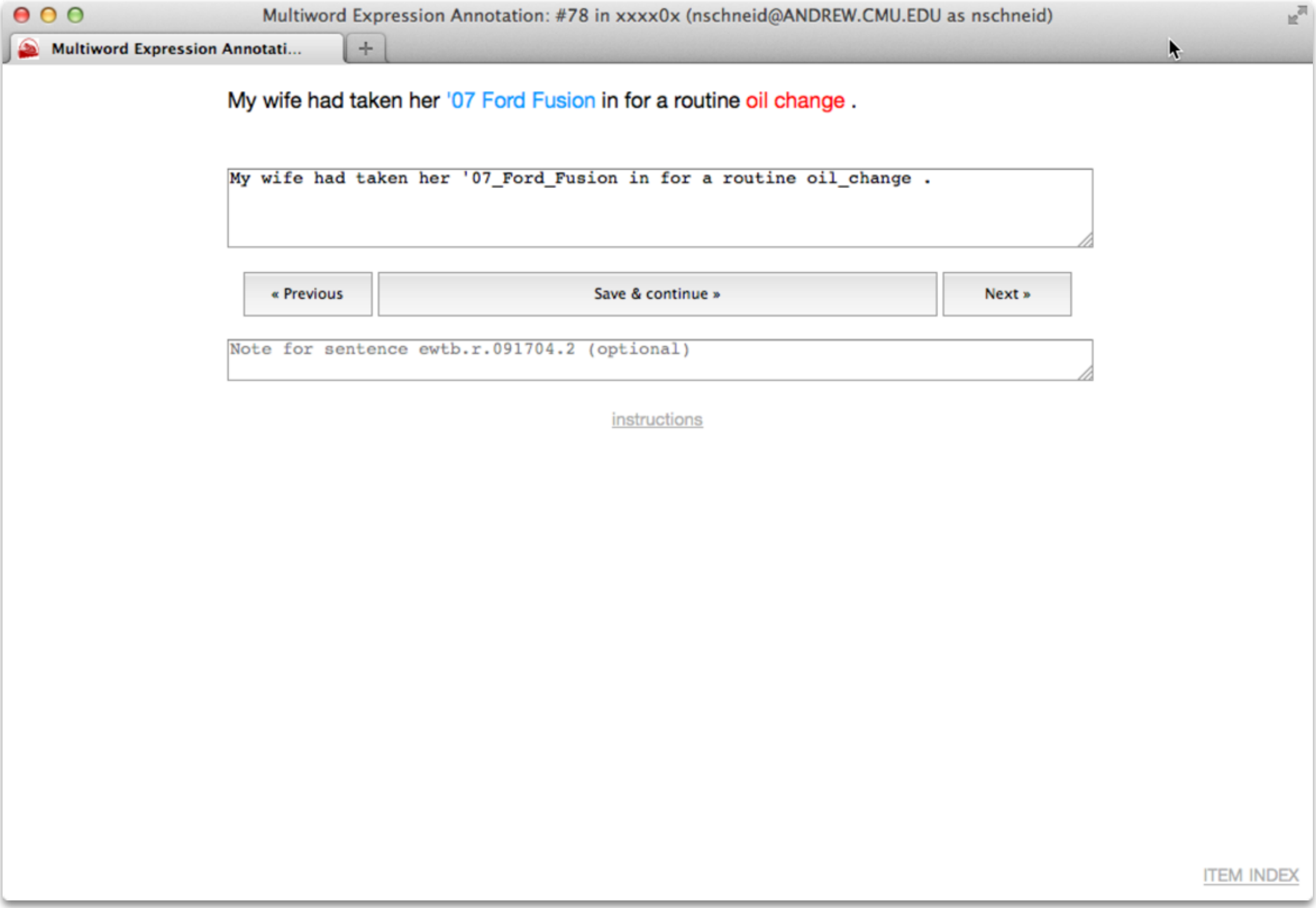
Examples

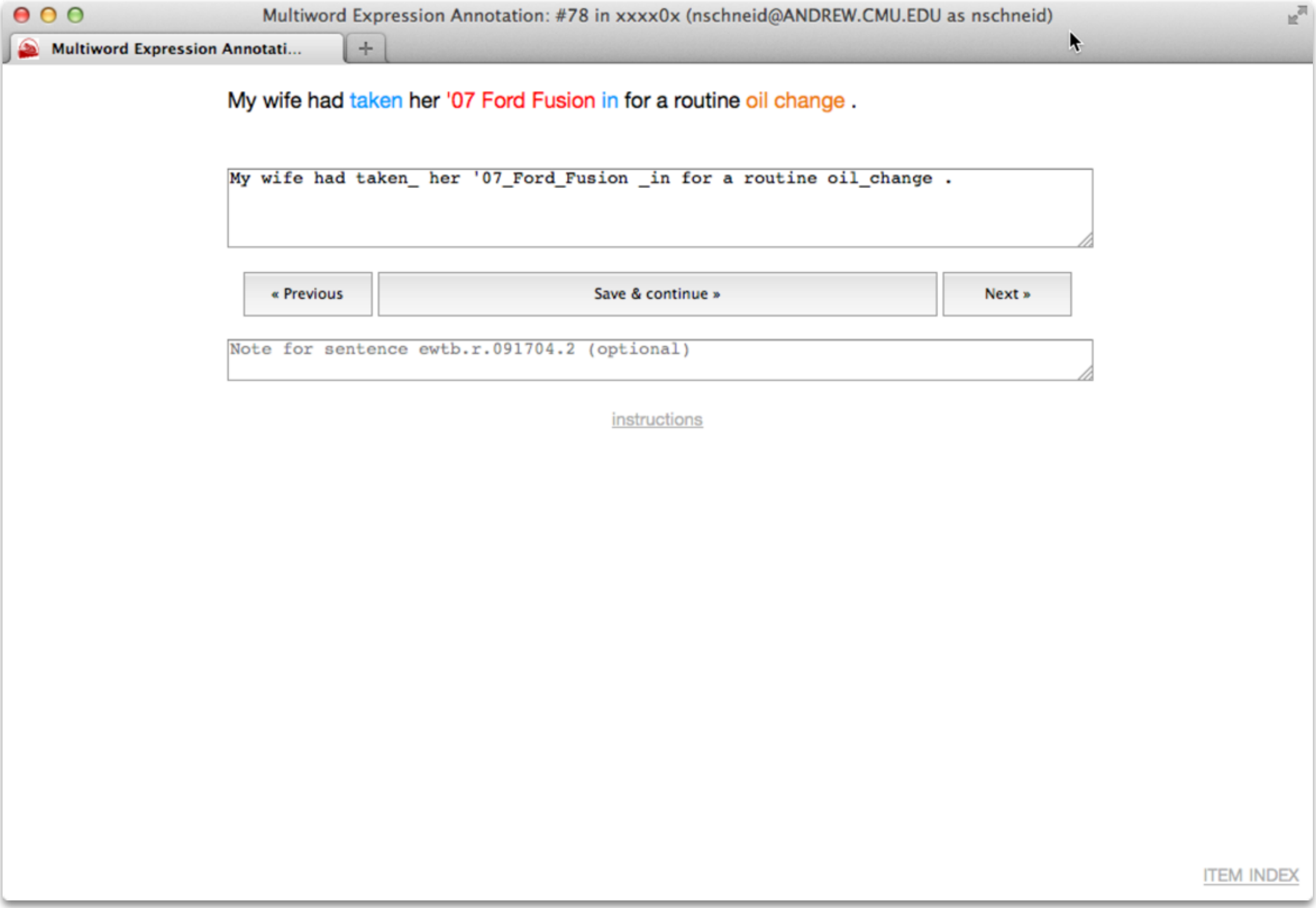
Among the animals that were available to touch were pony's , camels and EVEN AN OSTRICH !!!

No MWEs here. (This sentence is in the minority: 57% of all sentences/72% >10 words contain an MWE.)


Multiword Expression Annotation: #78 in xxxx0x (nschneid@ANDREW.CMU.EDU as nschneid)
Multiword Expression Annotati...
My wife had taken her '07 Ford Fusion in for a routine oil change .
My wife had taken her '07 Ford Fusion in for a routine oil change .
« Previous
Save & continue »
Next »
Note for sentence ewtb.r.091704.2 (optional)
instructions
ITEM INDEX


Multiword Expression Annotation: #78 in xxxx0x (nschneid@ANDREW.CMU.EDU as nschneid)
Multiword Expression Annotati...
My wife had taken her '07 Ford Fusion in for a routine oil change .
My wife had taken her '07_Ford_Fusion in for a routine oil change .
« Previous
Save & continue »
Next »
Note for sentence ewtb.r.091704.2 (optional)
instructions
ITEM INDEX


Multiword Expression Annotation: #78 in xxxx0x (nschneid@ANDREW.CMU.EDU as nschneid)
Multiword Expression Annotati...
My wife had taken her '07 Ford Fusion in for a routine oil change .
My wife had taken her '07_Ford_Fusion in for a routine oil_change .
« Previous
Save & continue »
Next »
Note for sentence ewtb.r.091704.2 (optional)
instructions
ITEM INDEX


Multiword Expression Annotation: #78 in xxxx0x (nschneid@ANDREW.CMU.EDU as nschneid)
Multiword Expression Annotati...
My wife had taken her '07 Ford Fusion in for a routine oil change .
My wife had taken_ her '07_Ford_Fusion _in for a routine oil_change .
« Previous
Save & continue »
Next »
Note for sentence ewtb.r.091704.2 (optional)
instructions
ITEM INDEX

The corpus

The corpus

- The entire **Reviews** subsection of the English Web Treebank (Bies et al. 2012), fully annotated for MWEs
  - 723 reviews
  - 3,800 sentences
  - 55,000 words
  - found 3,500 MWE instances

The corpus

- The entire **Reviews** subsection of the English Web Treebank (Bies et al. 2012), fully annotated for MWEs
  - 723 reviews
  - 3,800 sentences
  - 55,000 words
  - found 3,500 MWE instances
- Every sentence: negotiated consensus between at least 2 annotators
  - IAA between *pairs*: ~77%

Other English corpora

Other English corpora

- **SemCor** (Miller et al. 1993)
  - Lexical annotation with WordNet synsets
  - NEs, compound nominals, some phrasal verbs

Other English corpora

- **SemCor** (Miller et al. 1993)
  - Lexical annotation with WordNet synsets
  - NEs, compound nominals, some phrasal verbs
- **Prague CEDT** (Hajič et al. 2012)
  - NEs, light verb constructions, phrasal idioms, multiword `tlemmas`

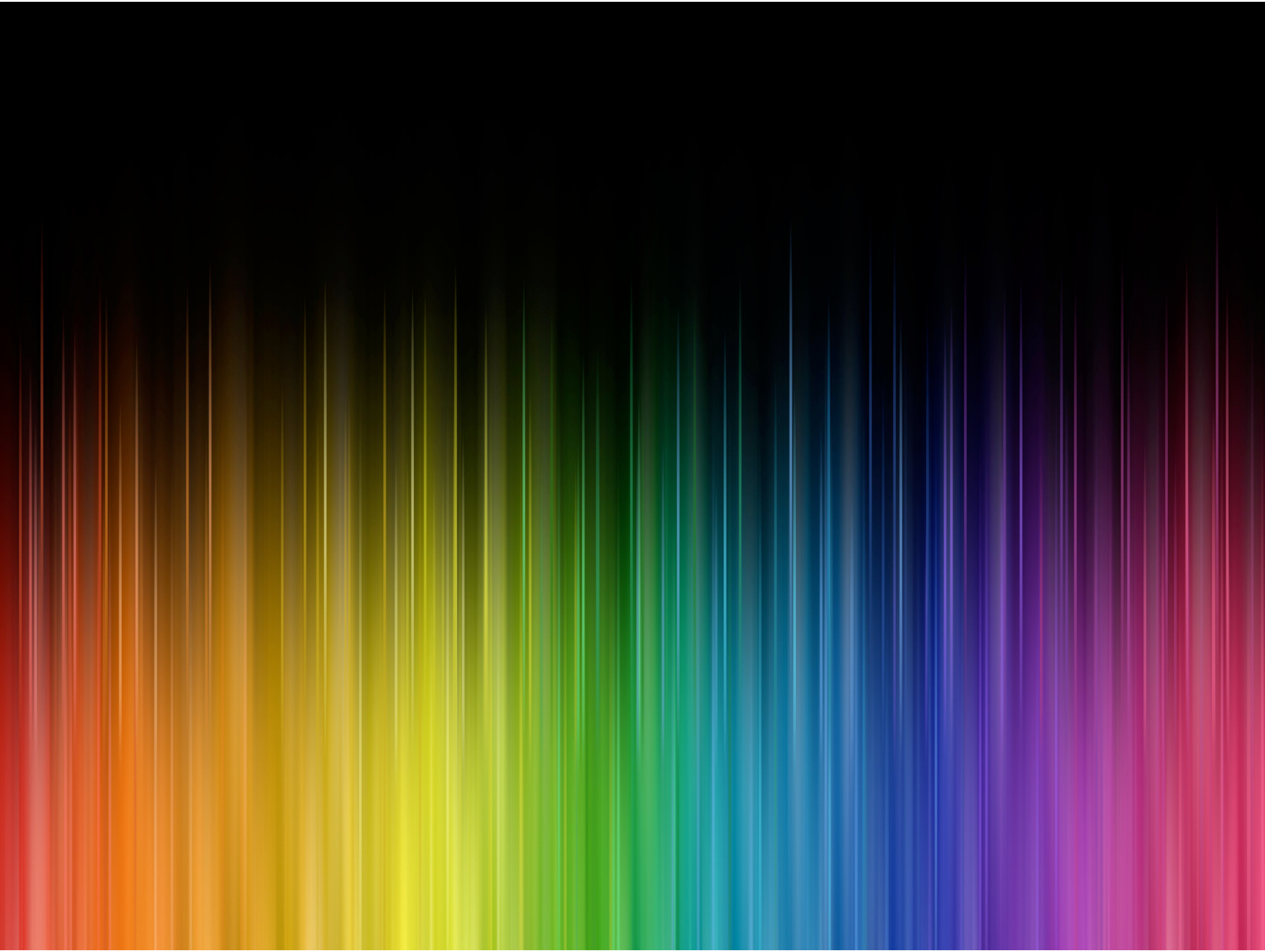
Other English corpora

- **SemCor** (Miller et al. 1993)
  - Lexical annotation with WordNet synsets
  - NEs, compound nominals, some phrasal verbs
- **Prague CEDT** (Hajič et al. 2012)
  - NEs, light verb constructions, phrasal idioms, multiword `tlemmas`
- **Wiki50** (Vincze et al. 2011)
  - NEs, compound nominals, LVCs, VPCs, phrasal idioms

CMWE, a text corpus comprehensively annotated with **multiword expressions**

- ✦ 55k words of English web reviews
- ✴ fully heterogeneous MWEs
- ✴ shallow groupings, allowing gaps
- ✴ strong vs. weak

tinyurl.com/mwecorpus



Preview:
A tool for identifying MWEs in context

(sequence tagging, but with gaps!)

Preview:
A tool for identifying MWEs in context

(sequence tagging, but with gaps!)

- **Discriminative lexical semantic segmentation with gaps: running the MWE gamut.** Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. *TACL* 2014.

CMWE, a text corpus comprehensively annotated with **multiword expressions**

- ✦ 55k words of English web reviews
- ✴ fully heterogeneous MWEs
- ✴ shallow groupings, allowing gaps
- ✴ strong vs. weak

tinyurl.com/mwecorpus

Many_thanks
(*Several thanks)

Thanks_a_million
(*Thanks a thousand)

Thanks_a_lot
(?Lots of thanks)

POS pattern	MWEs *contig.*	*gappy*	most frequent types (lowercased lemmas) and their counts
N_N	331	1	**customer service: 31** oil change: 9 wait staff: 5 garage door: 4
^_^	325	1	santa fe: 4 dr. shady: 4
V_P	217	44	**work with: 27** **deal with: 16** **look for: 12** **have to: 12** ask for: 8
V_T	149	42	**pick up: 15** **check out: 10** show up: 9 end up: 6 give up: 5
V_N	31	107	take time: 7 give chance: 5 waste time: 5 have experience: 5
A_N	133	3	front desk: 6 top notch: 6 last minute: 5
V_R	103	30	**come in: 12** come out: 8 take in: 7 stop in: 6 call back: 5
D_N	83	1	**a lot: 30** **a bit: 13** a couple: 9
P_N	67	8	**on time: 10** in town: 9 in fact: 7
R_R	72	1	**at least: 10** at best: 7 as well: 6 of course: 5 at all: 5
V_D_N	46	21	**take the time: 11** do a job: 8
V~N	7	56	*do job: 9* *waste time: 4*
^_^_^	63		home delivery service: 3 lake forest tots: 3
R~V	49		***highly recommend: 43*** *well spend: 1* *pleasantly surprise: 1*
P_D_N	33	6	over the phone: 4 on the side: 3 at this point: 2 on a budget: 2
A_P	39		pleased with: 7 happy with: 6 interested in: 5
P_P	39		**out of: 10** due to: 9 because of: 7
V_O	38		**thank you: 26** get it: 2 trust me: 2
V_V	8	30	get do: 8 let know: 5 have do: 4
N~N	34	1	*channel guide: 2* *drug seeker: 2* *room key: 1* *bus route: 1*
A~N	31		*hidden gem: 3* *great job: 2* *physical address: 2* *many thanks: 2* *great guy: 1*
V_N_P	16	15	**take care of: 14** have problem with: 5
N_V	18	10	mind blow: 2 test drive: 2 home make: 2
^_$	28		bj s: 2 fraiser 's: 2 ham s: 2 alan 's: 2 max 's: 2
D_A	28		**a few: 13** **a little: 11**