CAMBRIDGE
UNIVERSITY PRESS

Language, Gender and Sport

Insights from the Cambridge English Corpus
— October 2016

CONTENTS

Introduction

This report addresses the issue of gender representation in sport, and investigates how our language changes when we talk about women versus when we talk about men. For this research, the representation of gender in three contexts is explored and contrasted:

1. General English
2. English associated with sport
3. English associated specifically with the 2016 Olympics

The research makes use of corpus data collected by Cambridge University Press (see further information on the data below). The data was analysed using the web-based corpus software Sketch Engine: www.sketchengine.co.uk.

Methodologies used

The findings presented in this paper are grouped thematically, and are the result of an approach which sought to (i) validate hypotheses using the data, and (ii) allow ideas to be generated through an iterative, open-minded exploration of the data.

A combination of quantitative and qualitative methodologies were used, including (but not limited to) the following search-types:

- Frequency lists and key word lists.
- Word Sketches: a corpus based summary of a word's grammatical and collocational behaviour.
- Sketch Differences: these highlight the differences between the Word Sketches of two separate words, providing insight into the collocations and grammatical patterns which are shared, and those which are specific to each word.
- Concordance lines: generating a concordance line allows you to see every instance of a particular word or phrase in a given corpus. These results can then be queried and sorted to yield further insight. A combination of quantitative and qualitative approaches was taken in concordance analysis for this research.

About the data

The Cambridge English Corpus

The Cambridge English Corpus (CEC) is a multi-billion word collection of English language, containing both written and spoken English. Written data is drawn from a range of sources, such as newspapers, magazines, novels, letters, emails, textbooks, websites and many more. Spoken data is taken from everyday conversations, telephone calls, radio broadcasts, business meetings, presentations, speeches, and university lectures.

The Cambridge Sports Corpus

The Sports Corpus is a 150 million-word subset of the Cambridge English Corpus, containing only data which is tagged as being related to the subject category of *sport*.

The Cambridge Olympics Corpus

The Olympics Corpus is an 11.5 million-word corpus. The data in this corpus was drawn from the web over the course of the Rio Olympics 2016 using seed words and specific URLs to ensure maximum relevance.

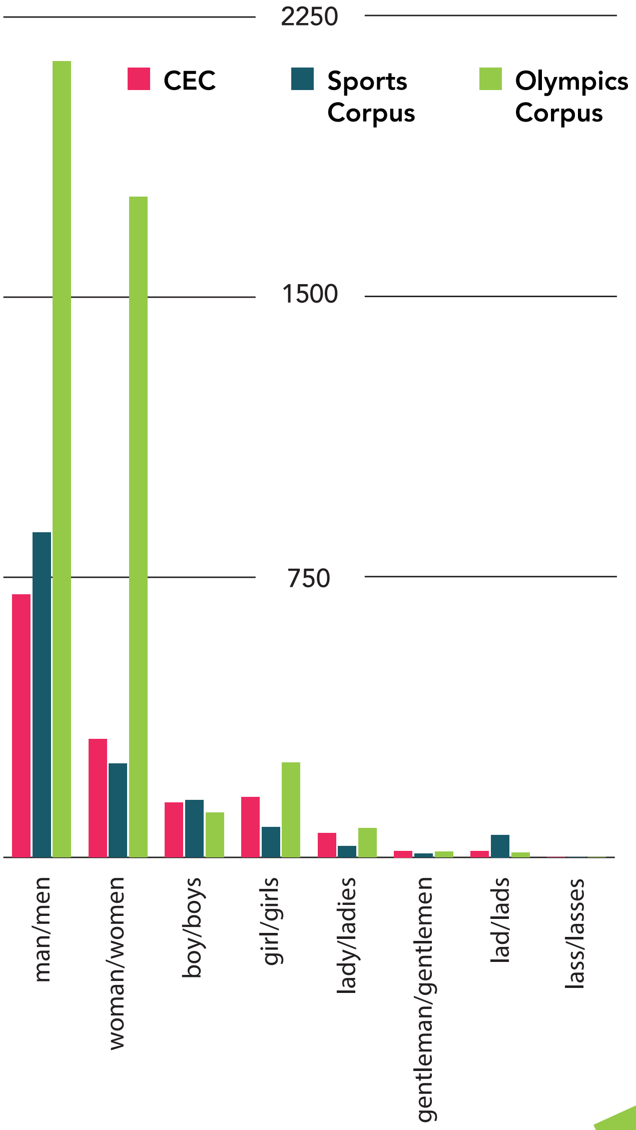
Findings

Note that the trends described regarding the Sports Corpus and the Olympics Corpus refer to men and women generally mentioned in these corpora, rather than specifically the athletes and sportspeople themselves; this could be fans of sport, spectators of sport, partners of sportspeople, or anyone else mentioned in the broad context of sport and the Olympics.

Women get less airtime in general, but especially in sport

The table and graph below shows the prevalence of lemmas related to men/women in (i) the Cambridge English Corpus, (ii) the Sports Corpus, and (iii) the Olympics Corpus. Numbers are normalised and refer to occurrences per million words:

	CEC	SPORTS CORPUS	OLYMPICS CORPUS
man/men	703.4	870.08	2,131.20
woman/ women	316.95	251.69	1,768.80
boy/boys	146.65	152.84	120.03
girl/girls	161.67	81.32	253.76
lady/ladies	64.5	30.33	77.98
gentleman/ gentlemen	17.42	10.17	15.61
lad/lads	16.65	60.29	13.09
lass/lasses	1.29	1.18	0.89



In all three corpora, we see a striking imbalance in the frequencies of the lemmas related to men and women.

In the Cambridge English Corpus, we see *man/men* more than twice as much as *woman/women*. We also see this imbalance with other gender pairs in the CEC, such as *boy/girl, lady/gentleman* and *lad/lass*.

Women's sports are often considered to be under-reported by the media, and this concern is validated by our Sports Corpus data; we often see an even greater gender imbalance than in the CEC, with more than three times as many mentions of man/men than woman/women.

But interestingly, the degree of imbalance for certain gender pairs is less marked in the Olympics Corpus where women seem to be catching up - the frequency gap with men appears to be significantly smaller in this context. However, this could also be influenced by a higher tendency for gender marking in women's sports – more about this later.

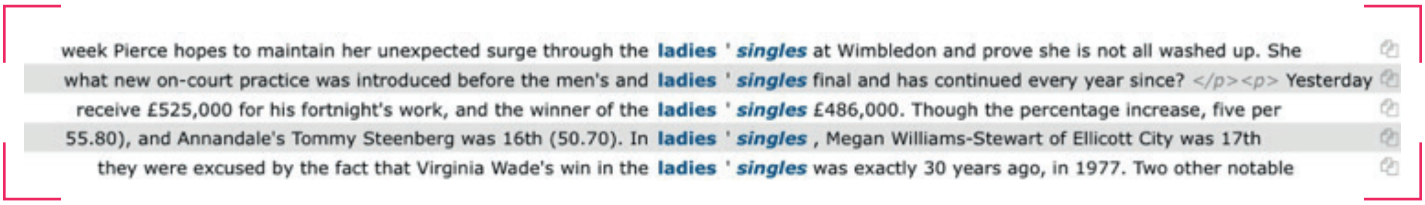
Note that for all of the words related to females, the relative frequency drops in the Sports Corpus when compared to the CEC; again, this is consistent with the idea that women's sports are under-represented. Note also that the words relating to men all go up, with the exception of *gentleman*. We see a particularly marked rise in the frequency of *lad/lads*, which is more negatively loaded in terms of behavioural connotations.

We see a striking increase in the relative frequencies of both *man/men* and *woman/women* in the Olympics Corpus when compared with the CEC and the Sports Corpus. This could be due to the use of these words as gender markers for sports. It could also be related to the extremely high degree of scrutiny Olympic athletes experience from the media; the Olympics is all about the Olympian, so we might expect to see an increase in the language used to refer to them when compared with the CEC.

Gender marking in sports: is the situation changing?

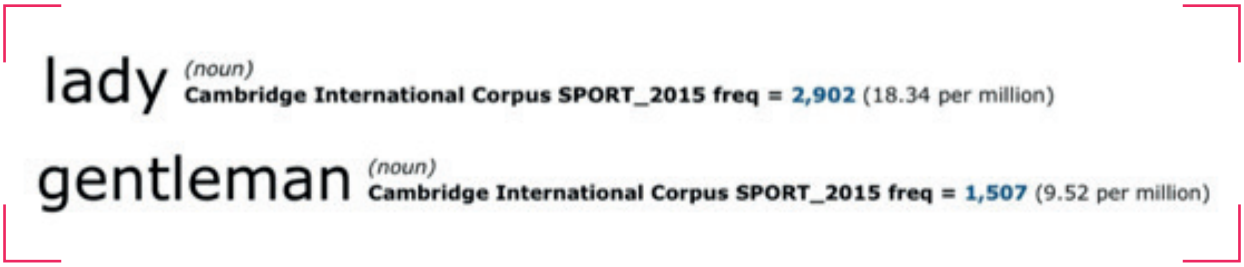
Overt gender marking is much more common for women's participation in sport, both in terms of the sport itself (*ladies' singles*) and the athletes participating (*woman golfer*). We do not see the same comparable gender marking tendency in the use of the word 'men'.

We also see instances of gender marking with *lady,* such as *ladies' singles*:



```
               week Pierce hopes to maintain her unexpected surge through the ladies ' singles at Wimbledon and prove she is not all washed up. She
          what new on-court practice was introduced before the men's and ladies ' singles final and has continued every year since? </p><p> Yesterday
           receive £525,000 for his fortnight's work, and the winner of the ladies ' singles £486,000. Though the percentage increase, five per
55.80), and Annandale's Tommy Steenberg was 16th (50.70). In ladies ' singles , Megan Williams-Stewart of Ellicott City was 17th
          they were excused by the fact that Virginia Wade's win in the ladies ' singles was exactly 30 years ago, in 1977. Two other notable
```

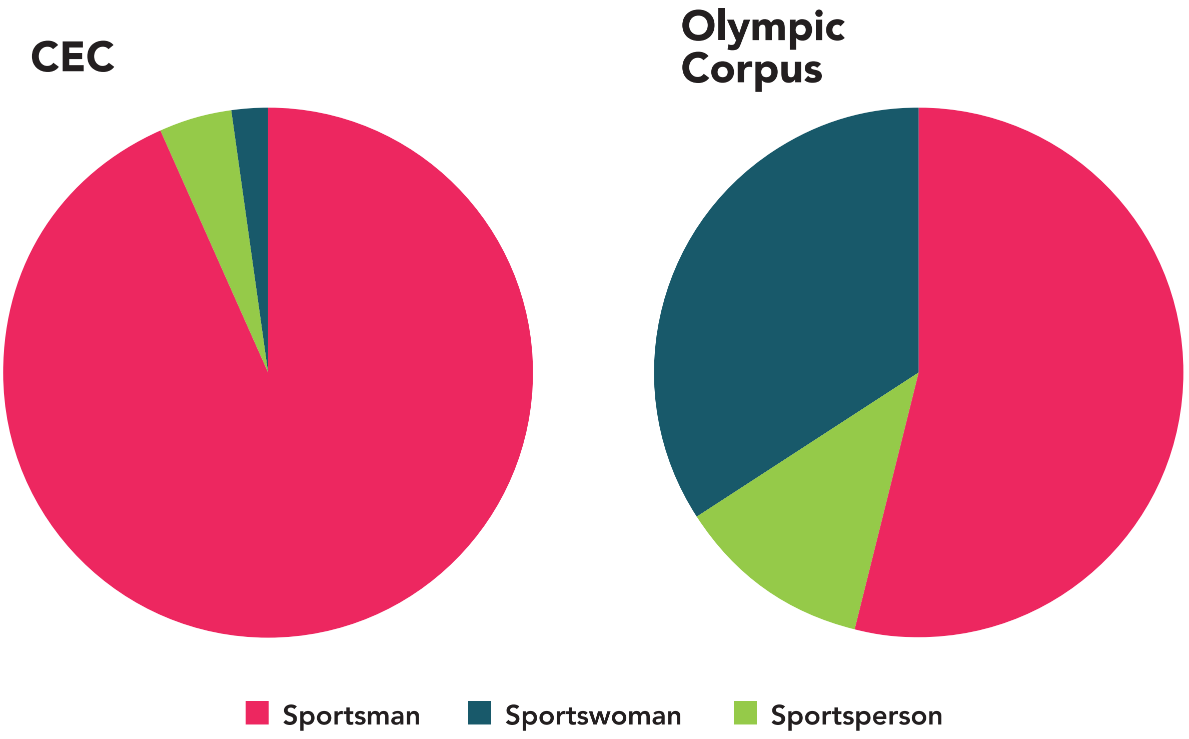
In fact, the word *lady* is twice as common in the Sports Corpus as its counterpart *gentleman*:



lady *(noun)* **Cambridge International Corpus SPORT_2015 freq = 2,902** (18.34 per million)

gentleman *(noun)* **Cambridge International Corpus SPORT_2015 freq = 1,507** (9.52 per million)

The charts below show the relative frequencies of the lemmas *sportsman, sportswoman* and the gender-neutral *sportsperson* in the Cambridge English Corpus and the Olympics Corpus:



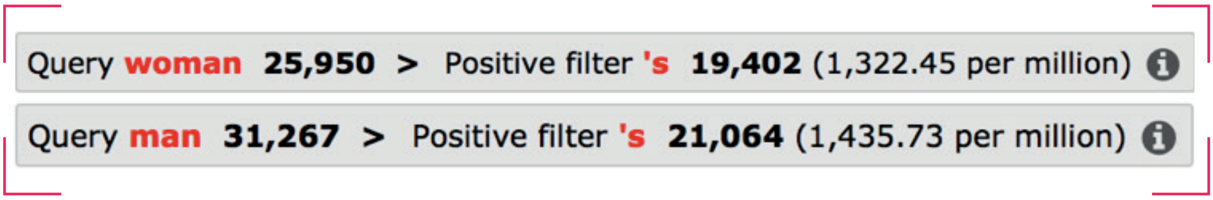
Though we see a higher prevalence of *sportsman* overall in both the CEC and the Olympics Corpus, note that the frequencies of *sportsman/sportswoman/sportsperson* are considerably more balanced in the Olympics Corpus – particularly *sportsperson*. The media has been the subject of harsh criticism over recent years for sexist reporting of men and women's sports; is the relatively high usage of the gender-neutral *sportsperson* an attempt to address this?

That said, we still see many more mentions of *female athletes* than *male athletes* in the Olympics Corpus. This may reflect the growing participation of women in Olympic sports as well as the tendency to mark women's sport more than men's.

It is interesting to note that there are 1,756 instances (118.57 per million) of *Olympian* in the Olympics corpus - a gender-neutral term. However, we see 45 instances of *female Olympian* and only two instances of *male Olympian*.

Women's vs Men's

Looking at the number of instances of *man* and *woman* followed by possessive *'s* affords some insight into the degree of gender marking for sports during the Olympics. In the Olympics corpus, we see the following frequencies of the lemmas *woman* and *man* followed by the possessive *'s*:

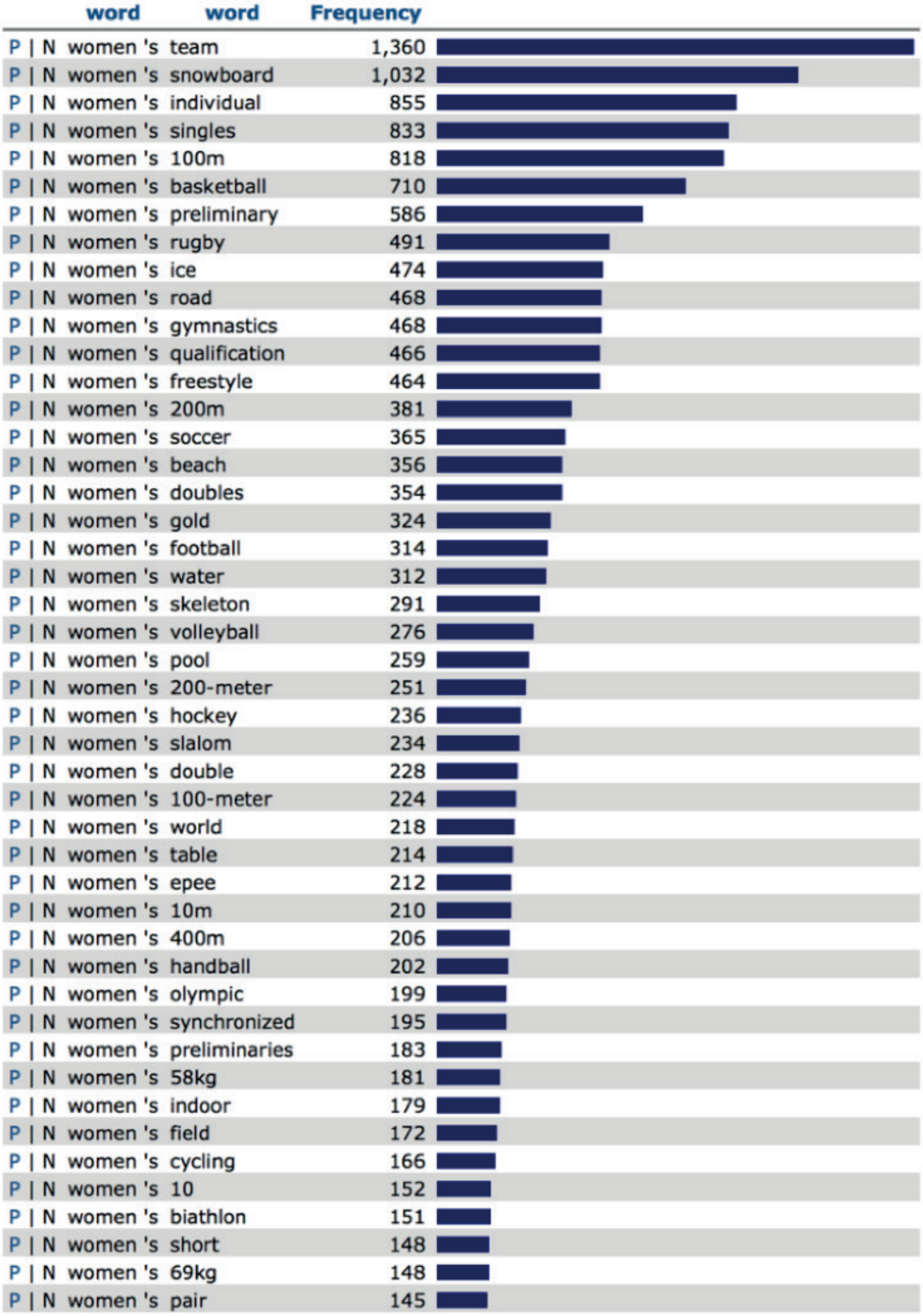


Query **woman** **25,950** > Positive filter **'s** **19,402** (1,322.45 per million)

Query **man** **31,267** > Positive filter **'s** **21,064** (1,435.73 per million)

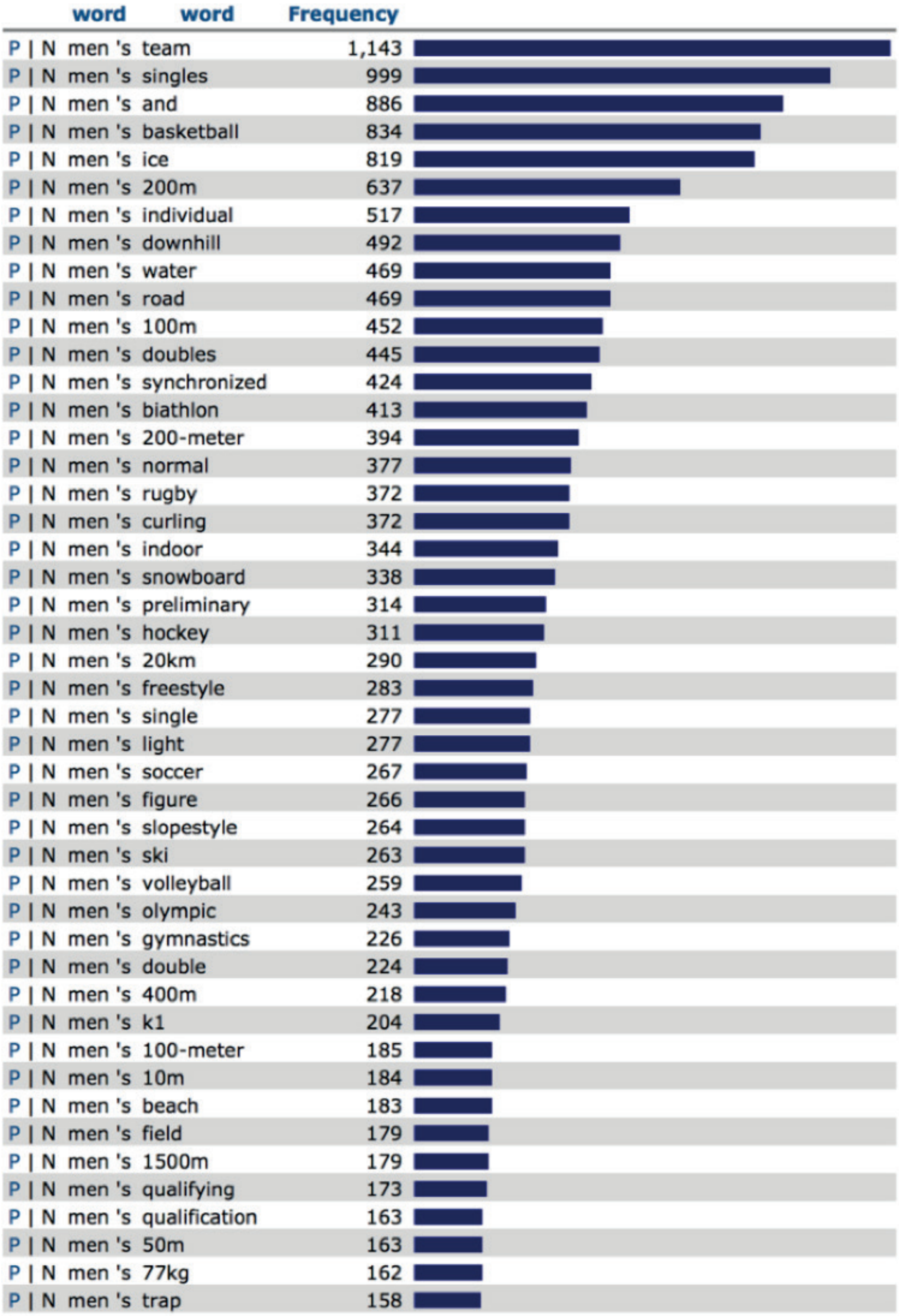
Though the raw number of possessives is higher for men, the proportions of possessives are not equal for men and women; 75% of the time, *woman* is followed by the possessive *'s*, and 67% of the time *man* is followed by the possessive 's. This points towards a slightly higher prevalence of gender marking for women's sports; although as the differences is fairly small, it would seem the situation might not be as it is often portrayed to be.

Women's is commonly followed by:



	word	word	Frequency
P | N	women 's	team	1,360
P | N	women 's	snowboard	1,032
P | N	women 's	individual	855
P | N	women 's	singles	833
P | N	women 's	100m	818
P | N	women 's	basketball	710
P | N	women 's	preliminary	586
P | N	women 's	rugby	491
P | N	women 's	ice	474
P | N	women 's	road	468
P | N	women 's	gymnastics	468
P | N	women 's	qualification	466
P | N	women 's	freestyle	464
P | N	women 's	200m	381
P | N	women 's	soccer	365
P | N	women 's	beach	356
P | N	women 's	doubles	354
P | N	women 's	gold	324
P | N	women 's	football	314
P | N	women 's	water	312
P | N	women 's	skeleton	291
P | N	women 's	volleyball	276
P | N	women 's	pool	259
P | N	women 's	200-meter	251
P | N	women 's	hockey	236
P | N	women 's	slalom	234
P | N	women 's	double	228
P | N	women 's	100-meter	224
P | N	women 's	world	218
P | N	women 's	table	214
P | N	women 's	epee	212
P | N	women 's	10m	210
P | N	women 's	400m	206
P | N	women 's	handball	202
P | N	women 's	olympic	199
P | N	women 's	synchronized	195
P | N	women 's	preliminaries	183
P | N	women 's	58kg	181
P | N	women 's	indoor	179
P | N	women 's	field	172
P | N	women 's	cycling	166
P | N	women 's	10	152
P | N	women 's	biathlon	151
P | N	women 's	short	148
P | N	women 's	69kg	148
P | N	women 's	pair	145

Men's is commonly followed by:



	word	word	Frequency
P | N	men 's	team	1,143
P | N	men 's	singles	999
P | N	men 's	and	886
P | N	men 's	basketball	834
P | N	men 's	ice	819
P | N	men 's	200m	637
P | N	men 's	individual	517
P | N	men 's	downhill	492
P | N	men 's	water	469
P | N	men 's	road	469
P | N	men 's	100m	452
P | N	men 's	doubles	445
P | N	men 's	synchronized	424
P | N	men 's	biathlon	413
P | N	men 's	200-meter	394
P | N	men 's	normal	377
P | N	men 's	rugby	372
P | N	men 's	curling	372
P | N	men 's	indoor	344
P | N	men 's	snowboard	338
P | N	men 's	preliminary	314
P | N	men 's	hockey	311
P | N	men 's	20km	290
P | N	men 's	freestyle	283
P | N	men 's	single	277
P | N	men 's	light	277
P | N	men 's	soccer	267
P | N	men 's	figure	266
P | N	men 's	slopestyle	264
P | N	men 's	ski	263
P | N	men 's	volleyball	259
P | N	men 's	olympic	243
P | N	men 's	gymnastics	226
P | N	men 's	double	224
P | N	men 's	400m	218
P | N	men 's	k1	204
P | N	men 's	100-meter	185
P | N	men 's	10m	184
P | N	men 's	beach	183
P | N	men 's	field	179
P | N	men 's	1500m	179
P | N	men 's	qualifying	173
P | N	men 's	qualification	163
P | N	men 's	50m	163
P | N	men 's	77kg	162
P | N	men 's	trap	158

It would seem from these lists that gender marking for some Olympic sports is much more common than others: *football, snowboarding, cycling, rugby* and swimming are all far more likely to be marked for *women* than *men*.

Gendered words with semantic prosody

Certain terms associated with men and women have particularly negative or positive connotations; we get an insight into these connotations through the collocations of each term, particularly those in modifier position. Some example terms and their modifiers from the Cambridge English Corpus can be seen below, listed in order of strength of collocation:

Feminine terms:

- Bimbo: *empty headed, brainless, blonde*
- Chick: *hippy/hippie, biker, skater, groovy, dixie, wacky, skinny, rock*
- Babe: *busty, bikini-clad, scantily-clad*
- Girl: *teenage, little, young, pretty, beautiful, lovely, poor*
- Lady: *elderly, old, young, lovely, fat, middle-aged, nice, naked, pretty, little*
- Lass: *bonny, comely, buxom, we, leggy, strapping*

Masculine terms:

- Bloke: *bald, decent, ordinary, burly, fat, middle-aged, good-looking, posh, nice, lovely*
- Guy: *nice, bad, tough, smart, go-to, fall, regular, little*
- Chap: *cheeky, likeable, affable, jolly, decent*
- Dude: *surfer, gnarly, cool*
- Fellow: *jolly, poor, amiable, handsome, distinguished*
- Lad: *we, young, working-class, strapping, smashing, lovely, brave*
- Gent: *portly, dapper, elderly*

Feminine appearance and masculine behaviour

Although we see more mentions of *men* than *women* in the corpus, *feminine* gets almost twice as many mentions in the Cambridge English Corpus as *masculine*, with 4.44 /million and 2.94/million respectively. *Manly* (1.13/million), however, is more common than its counterpart *womanly* (0.35/million).

Ladylike has a relative frequency of 0.37/million in the Cambridge English Corpus, whereas the masculine counterpart *gentlemanly* occurs at 0.64/million. Alongside *ladylike*, we see collocations related to image, such as *daintiness* and *glamour*. Whereas *gentlemanly* commonly collocates with words that describe a certain kind of behaviour: *restraint, conduct, demeanour* and *behaviour*.

Girly has a relative frequency of 0.61/million in the Cambridge English Corpus. The kind of things we describe as *girly* include *giggle* and *chat*. We also see *girly* commonly modified by *too* – perhaps indicating concern that something might be excessively *girly*.

Effeminate (0.40/million) is also commonly modified by adverbs of quantity, such as (in order of salience) *vaguely, slightly, overly, somewhat* and *rather*. *Outrageously* and *wilfully* are also strong collocates of *effeminate*. It seems acceptable to be *girly* or *effeminate* up to a certain point, but as with *girly*, there is significant concern with the degree…

In reporting and discussion of women's sport, there is often criticism for a heavy focus on the aesthetic rather than athleticism, or the sporting performance itself. We see evidence of this in the Sports Corpus, where we find *women* collocating with *clad*, as in *scantily clad*, and we also see a collocation with the verb *dress*.

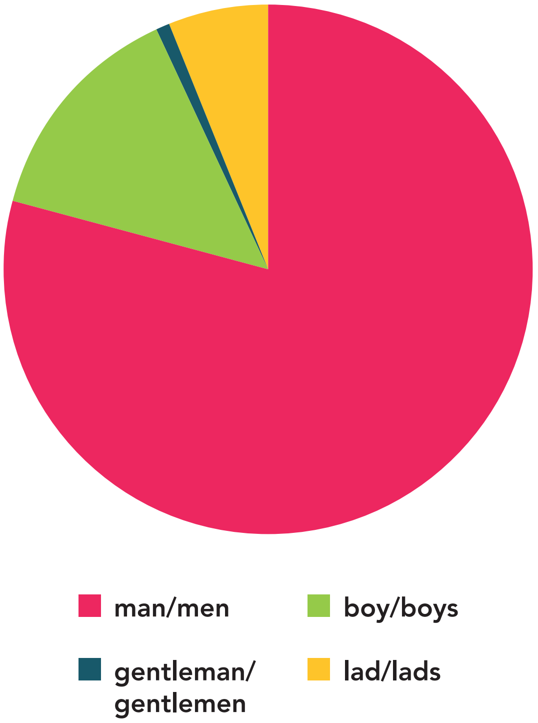
During the Olympics, the media was criticised for its obsession with the appearance and attire of female athletes. Again, we see evidence of this in the Olympics Corpus, where a Word Sketch for *women* shows a strong collocation (7.75) with the verb *wear*. We don't see this collocation with *men*.

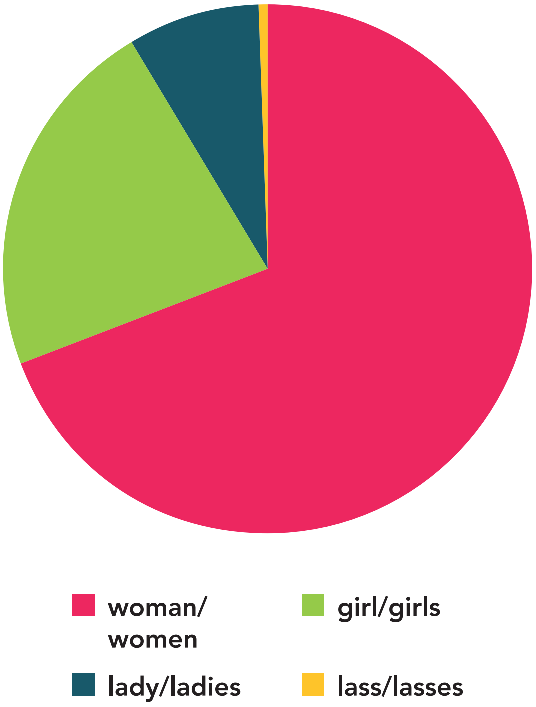
In the Sports Corpus, *married* and *unmarried* both make the list of top collocations for *women*, but not *men*. We're also much more interested in how old they are; *aged* is a top collocation for *women*, but not *men*.

Men are often said to be more competitive than women; we do see some evidence to support this in the Sports Corpus.

We see *men/man* collocating in subject position with verbs such as *mastermind, beat, win, dominate* and *battle*. Meanwhile, we see *woman/women* collocating in subject position with verbs such as *compete, participate* and *strive*.

Trivialisation and infantilisation





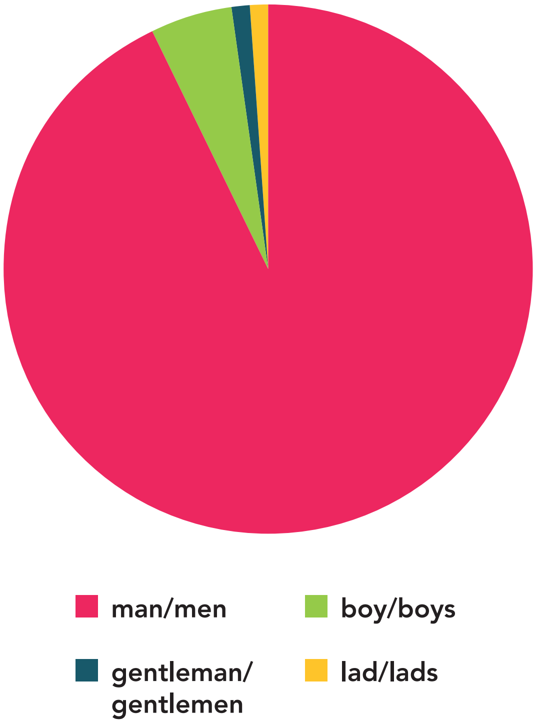
The charts above illustrate the breakdown of word-pairs used to refer to women when compared to men in the Sports Corpus. Note that this is not an exhaustive list of words used to refer to men/women, but includes some of the most common lemma pairs: *man vs woman, boy vs girl, lady vs gentleman, lad vs lass*. The charts compare the percentage use of these lemma pairs.

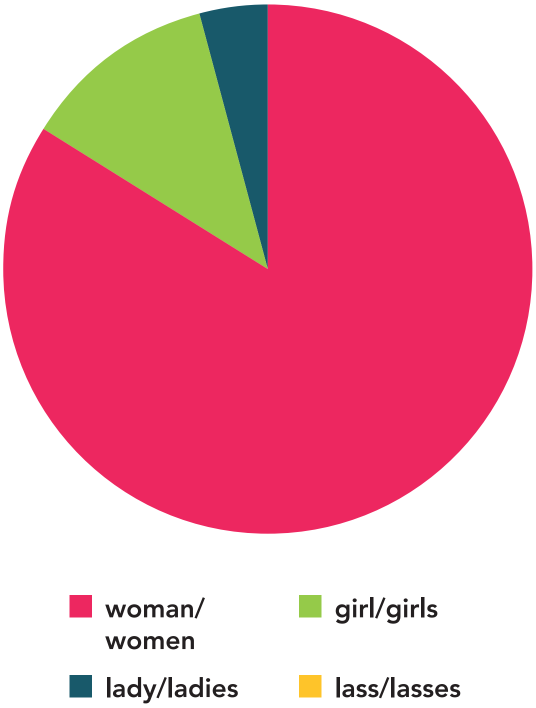
Note that in the Sports Corpus, the percentage for *girl* is higher than *boy*, and *lady* is higher than *gentleman*. It could be argued that the terms used to refer to women are more often characterised by either trivialisation/infantilisation, or feminisation, whereas a higher percentage of the time, men are referred to with the more neutral term *men*.

It may be that in the context of sports reporting, there is an attempt to reconcile the clash between traditional notions of femininity with the stereotypical sporty persona, which is more aligned with traditional notions of masculinity. Perhaps sports reporting seeks to balance this discrepancy by overemphasis on the more feminine/semantically loaded terms *girl* and *lady*.

In contrast, we see a proportion (6%) of references to men as *lads*, which has more negative semantic prosody than the alternatives.

Similarly, the charts below illustrate the breakdown of word-pairs used to refer to women and men in the Olympics Corpus:





We do see a striking increase in the relative frequency of *girl/girls* in the Olympics Corpus (253.76/million) when compared with the CEC (161.67/million) – the lemma is over 55% more common in the Olympics Corpus. This is of particular interest because we saw a decrease of *girl/girls* when comparing the CEC with the Sports Corpus (81.32/million).

We do not see the same pattern with *boy/boys*; in fact, we see the opposite, with a decrease in relative frequency in the Olympics Corpus when compared with both the Cambridge English Corpus and the Sports Corpus.

This supports the idea that there is a general tendency to infantilise women in sport more than men. However, the proportion of usage of *girl/girls* in the Olympics Corpus is less than in the Sports Corpus - 12% and 22% respectively. As we have speculated with the patterns in gender-marking, perhaps we are seeing a more balanced and neutral reporting of women's and men's sports by the media.

Interestingly, we don't see as many mentions of the semantically-loaded *lad/lads* as in the CEC or the Sports Corpus. What does this tell us about the culture of the Olympics by comparison with the culture of sport in general?

We generalise about women differently

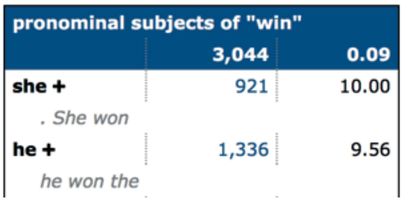
In the Cambridge English Corpus as a whole, it is much more common to find the adjectives *most* and *many* collocating with *women* (salience scores of 7.3 for both) rather than *men* (salience scores of 6.6 for *most men* and 6.3 for *many men*); we also see a stronger collocation with the verb *tend* for *women* than *men* (6.6 vs 5.5). We see these same trends in the Sports Corpus.

It could be argued that this indicates a tendency to generalise more about women; alternatively, it could be that we hedge these generalisations about women using qualifiers such as *most* and *many*, or verbs such as *tend*. Why might this be the case? Are we more reticent to treat women as a homogenous group?

Women clinch titles, men claim theirs

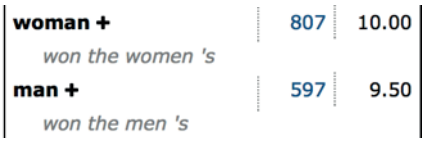
We are marginally more likely to talk about women *winning* than men:

Pronouns:



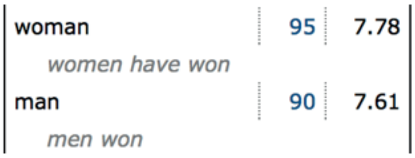
pronominal subjects of "win"		
	3,044	0.09
she +	921	10.00
. She won		
he +	1,336	9.56
he won the		

In object position:



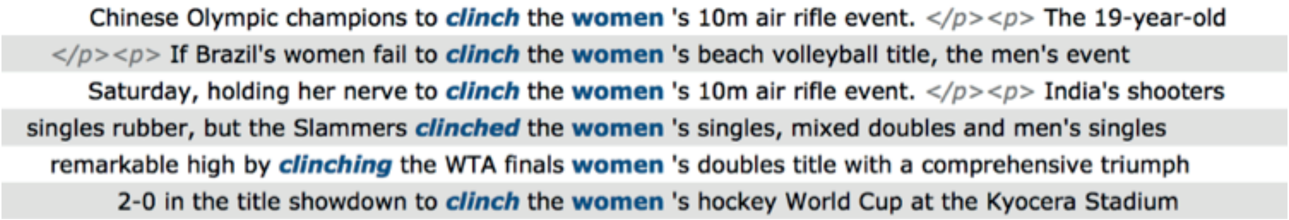
woman +	807	10.00
won the women 's		
man +	597	9.50
won the men 's		

In subject position:



woman	95	7.78
women have won		
man	90	7.61
men won		

However, the interesting thing is *how* women's titles are won. Women are more likely to *clinch* their titles, whereas men are likely to *claim* theirs:



Chinese Olympic champions to ***clinch*** the **women** 's 10m air rifle event. </p><p> The 19-year-old

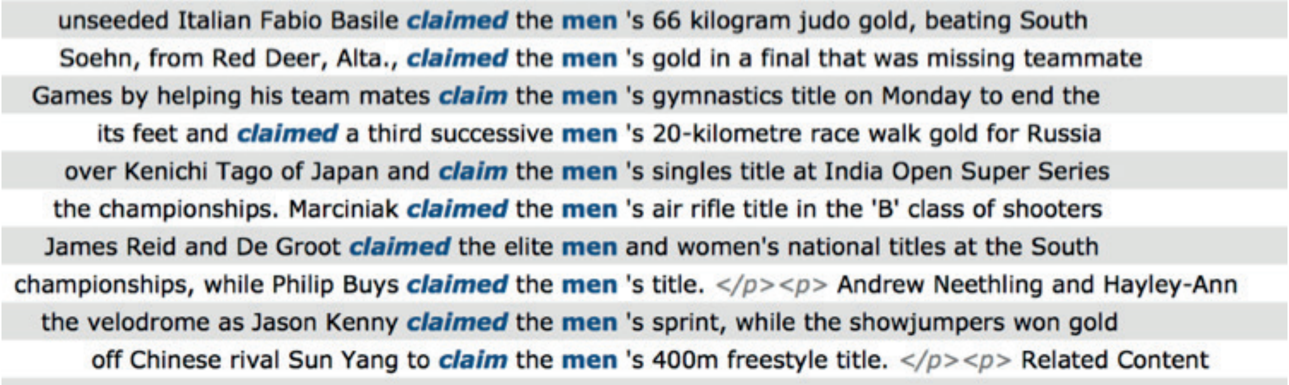
</p><p> If Brazil's women fail to ***clinch*** the **women** 's beach volleyball title, the men's event

Saturday, holding her nerve to ***clinch*** the **women** 's 10m air rifle event. </p><p> India's shooters

singles rubber, but the Slammers ***clinched*** the **women** 's singles, mixed doubles and men's singles

remarkable high by ***clinching*** the WTA finals **women** 's doubles title with a comprehensive triumph

2-0 in the title showdown to ***clinch*** the **women** 's hockey World Cup at the Kyocera Stadium



unseeded Italian Fabio Basile ***claimed*** the **men** 's 66 kilogram judo gold, beating South

Soehn, from Red Deer, Alta., ***claimed*** the **men** 's gold in a final that was missing teammate

Games by helping his team mates ***claim*** the **men** 's gymnastics title on Monday to end the

its feet and ***claimed*** a third successive **men** 's 20-kilometre race walk gold for Russia

over Kenichi Tago of Japan and ***claim*** the **men** 's singles title at India Open Super Series

the championships. Marciniak ***claimed*** the **men** 's air rifle title in the 'B' class of shooters

James Reid and De Groot ***claimed*** the elite **men** and women's national titles at the South

championships, while Philip Buys ***claimed*** the **men** 's title. </p><p> Andrew Neethling and Hayley-Ann

the velodrome as Jason Kenny ***claimed*** the **men** 's sprint, while the showjumpers won gold

off Chinese rival Sun Yang to ***claim*** the **men** 's 400m freestyle title. </p><p> Related Content

During the Olympics, the media was criticised for attributing the success of female athletes to others - particularly male partners or coaches; sure enough, we see *women* as the object of the verb *help* with a collocation salience of 7.31. We don't see this as a top collocation for *men*.

Conclusions

Through this analysis of millions of words from news and social media commentary around the 2016 Olympics, as well as more broadly in the domain of sport and general English, we have uncovered evidence which validates many concerns about the disparity in the representation of men and women in sport; the amount of airtime received by men and women remains unbalanced, the focus of attention on the aesthetic rather than the athletic for women remains a prevalent issue, the gender-marking of women's sports as the lesser 'other', and the trivialisation and infantilisation of women's sports are just some of the ways that gender inequality manifests linguistically.

However, it is hugely encouraging to see from analysis of the Olympics Corpus that the situation may be improving; the gap in the amount of airtime received by men and women is narrowing, and we see a higher prevalence of gender-neutral terms like *sportsperson* in our Olympics data (rather than the gender-marked counterparts). Though there is still a long way to go, it is reassuring to see linguistic evidence for the positive impact the Olympics can have on the representation of women in sport.

To cite this paper:

Cambridge University Press (2016). *Language, Gender and Sport.* Part of the Cambridge Papers in ELT series. [pdf] Cambridge: Cambridge University Press.

Available at cambridge.org/betterlearing