A quantitative analysis of gender differences in movies using psycholinguistic normatives (Ramkrishna, A, et al)

Psycho-linguistic.
Uses a metric known as gender ‘ladeness’.

Based on a set of 925 manually selected terms with their feminine/masculine scores(e.g ‘infantry’ has a higher masculine score than ‘gorgeous’). This set of words is expanded using lexicon expansion. The score of a sentence is simply the average score of all the words in the sentence.

To expand the lexicon, the authors did a large scale Yahoo! Search on all English words, collecting the top 500 results for each word then the top 10,000 words. There may be need to develop an alternative model to handle this model expansion.

Gender-Distinguishing Features in Film Dialogue (Schofield, A, et al)

Made use of NLTK. This model was built on the following sets of features:

1.) Unigrams, bigrams and trigrams. N-grams with a frequency of less than 5 were discarded
2.) VADER sentiment scores (positive, negative, composite)
3.)V/A/D (Valence, arousal and dominance) scores
3.) Average tokens (words) per line.
4.) Average token length

The paper reports good accuracy (~67%) using just the unigrams (words like ‘he’ and ‘him’, for example are favoured by female speakers while ‘she’ and ‘her’ are favoured by male speakers). Including the other features (bi-/trigrams, vader v/a/d features improved results only marginally (to 72%).

Do women and men really live in different cultures? Evidence from the BNC (Schmid, H)

Discusses the Leech-Fallon ‘difference coefficient’ (a method of distinguishing words spoken by males vis-a-vis those spoken by females. The coefficient has a range of -1 (negative scores mean that a word occurs more frequently in utterances attributed to women, positive ones that it is more often used in male utterances).

Reviews language more holistically. Considers three aspects that distinguish male/female conversation:

1.) Conversational behaviour: 'women's words', hesitation and hedges, minimal responses, questions
2.) Domains with expected female preponderance: clothing, colours, home, food and drink, body and health, personal reference, personal relationships, temporal deixis
3.)Domains with expected male preponderance: swearwords

In general from the literature, the following will be key in developing the model:

1.) The Leech-Fallon ‘difference coefficient’
2.) N-grams will be useful
3.) It may help to think of language ‘domains’ (e.g ‘clothing’, ‘sports’) and how popular they are with each gender.