**Unit 5**

**NLP Applications (Continued) :**

**(Disclaimer: All content is from various internet sources and for academic purpose only)**

Machine translation - Basic issues in MT.

There are several challenges that language translation poses for machine translation systems:

1. Ambiguity: Language is often ambiguous, and the same word or phrase can have multiple meanings depending on the context in which it is used. This can make it difficult for machine translation systems to accurately translate words and phrases.
2. Idioms and colloquialisms: Idioms and colloquialisms are expressions that are unique to a particular language or culture, and they do not have a direct equivalent in other languages. This can make it difficult for machine translation systems to accurately translate these expressions.
3. Linguistic diversity: Machine translation also has to deal with the diversity and complexity of natural languages, which can vary in terms of syntax, morphology, semantics, pragmatics, and style. For example, translating between languages that have different word orders, such as English and Japanese, may require reordering the words or phrases to preserve the meaning and coherence.
4. Cultural differences: Language is closely tied to culture, and different cultures have different ways of expressing ideas and concepts. This can make it difficult for machine translation systems to accurately translate cultural references and concepts.
5. Data quality: Machine translation relies on large amounts of parallel data, which are pairs of sentences in different languages that have the same meaning. However, parallel data can be scarce, noisy, or biased, depending on the source and domain. For example, translating between low-resource languages, such as Kannada, Telugu or Urdu may require collecting and cleaning data from various sources, such as web pages, subtitles, or dictionaries. Moreover, parallel data may reflect cultural or political biases, such as gender stereotypes, propaganda, or censorship, which can affect the accuracy and fairness of machine translation.
6. Language morphology: We may find more spelling variations in morphological rich language. - In morphological rich language, multiple words are combined together, which effects the word alignment in parallel corpus
7. Word sense disambiguation(WSD)- Sometimes one word can have multiple meanings. So, it creates difficulty in aligning the source word with the target word.
8. Evaluation metrics : Machine translation also faces the challenge of evaluating the quality and usefulness of the output, which can be subjective and dependent on various factors, such as the task, the domain, the user, and the reference. For example, measuring the similarity between the output and a human reference translation using metrics such as BLEU or METEOR may not capture the adequacy or fluency of the output, or the diversity or creativity of the translation.
9. The language can be divided into 6 categories based on the subject, object and verb order in the language. Among them, 3 categories are most popular

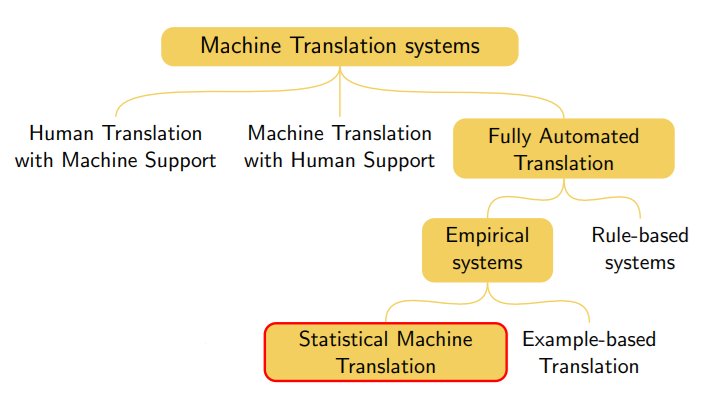
Subject-Object-Verb(SOV) language -Indian languages

Subject-Verb-Object(SVO) language- English & European lang

VSO language- Hebrew, Arabic

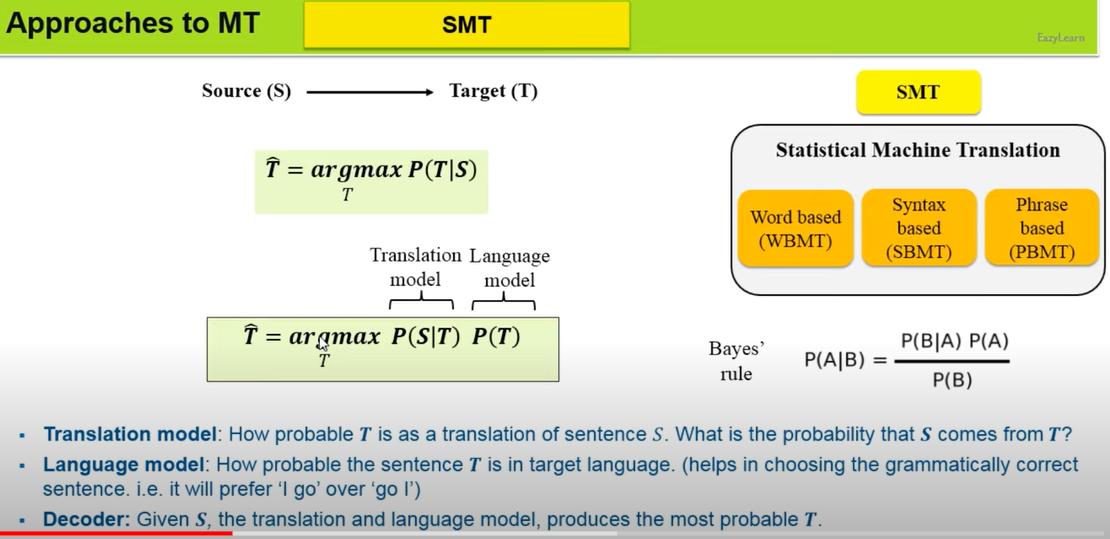
Statistical translation faces difficulty in translating the language from one language family to another language family. We need to perform long distance reordering for translation between different families of language. Syntax based translation is used for capturing word order.

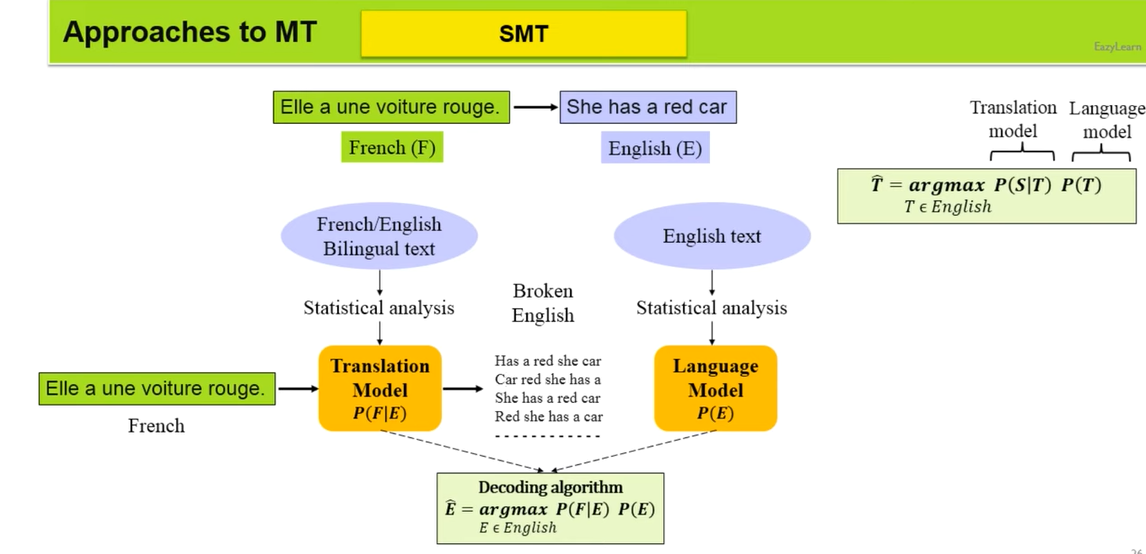
**Statistical translation**

****

**Introduction**

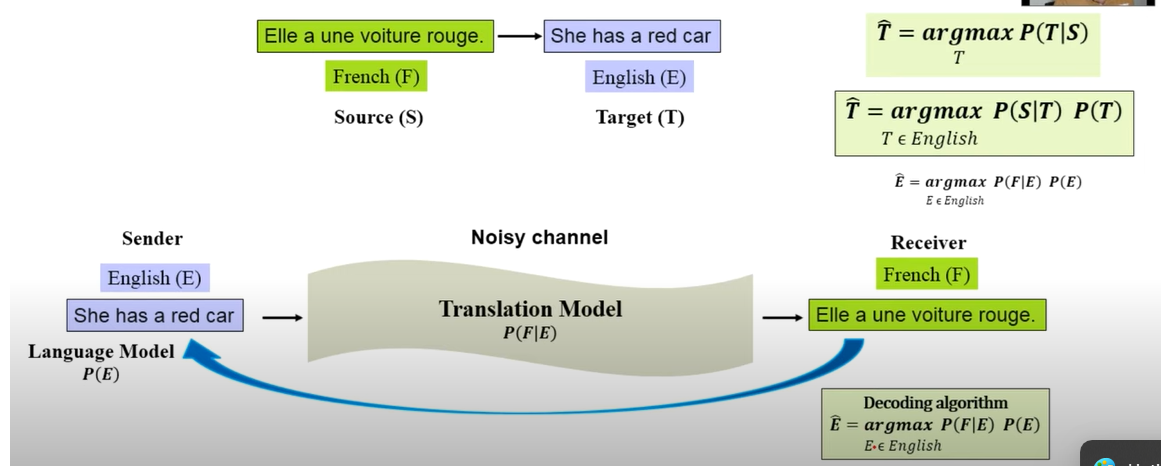
* Parallel corpora are available in several language pairs
* Basic idea is to use parallel corpus as a training set of translation examples





**The Noisy Channel Model**

* Goal: Translation system from French to English or Tamil to Hindi (Source to target language)

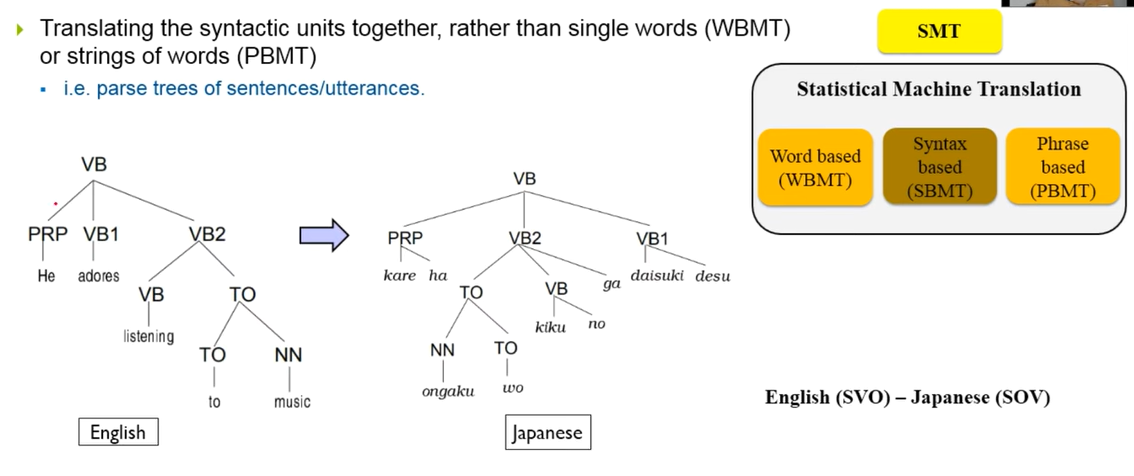
****

**Types of SMT**

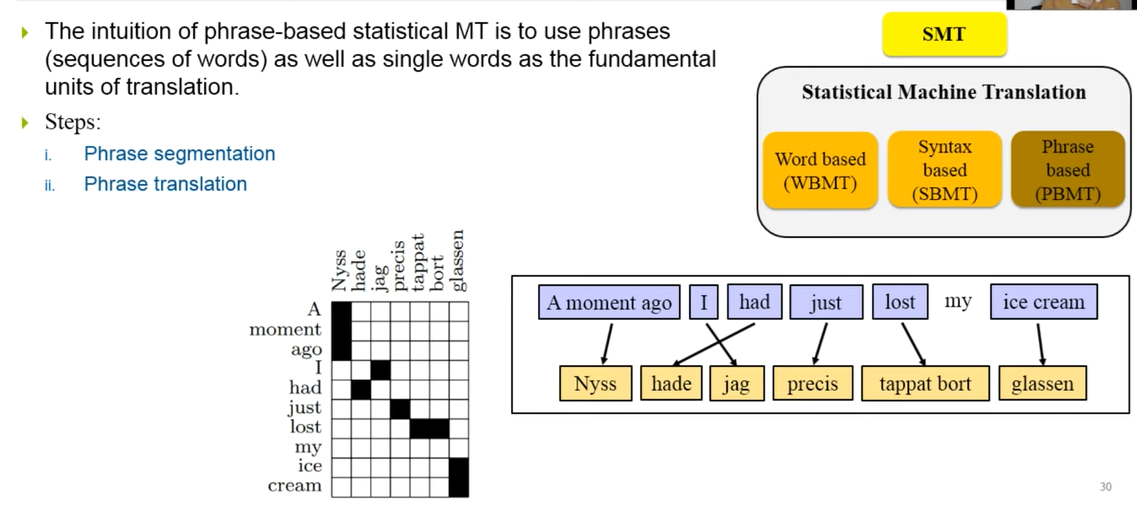
**WBMT:**

****

**SBMT:**

****

**PBMT:**

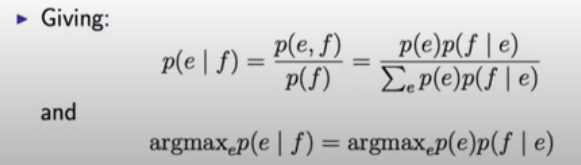
****

**Noise based translation**

* Have a model p(e|f) which estimates conditional probability of any English sentence e given the French sentence f. Use the training corpus to set the parameters.
* A noisy channel model has two components

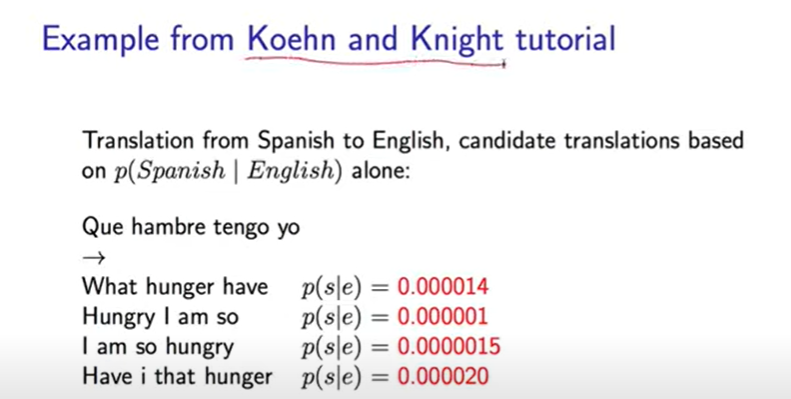
P(e) – the language model

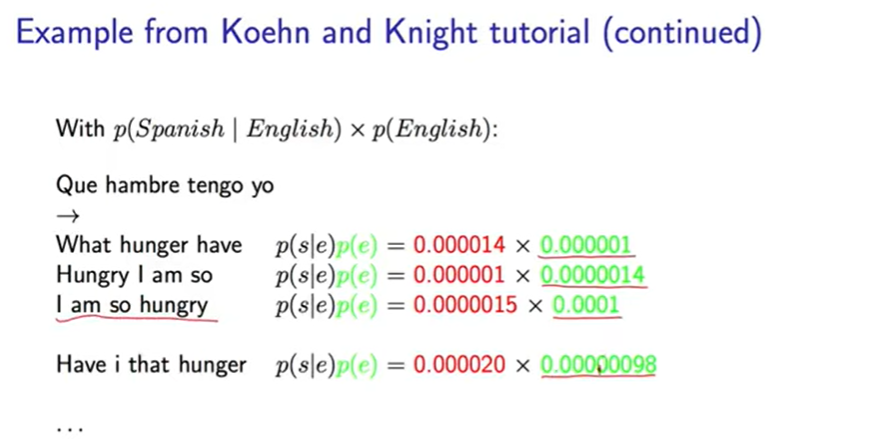
P(f|e) the translation model

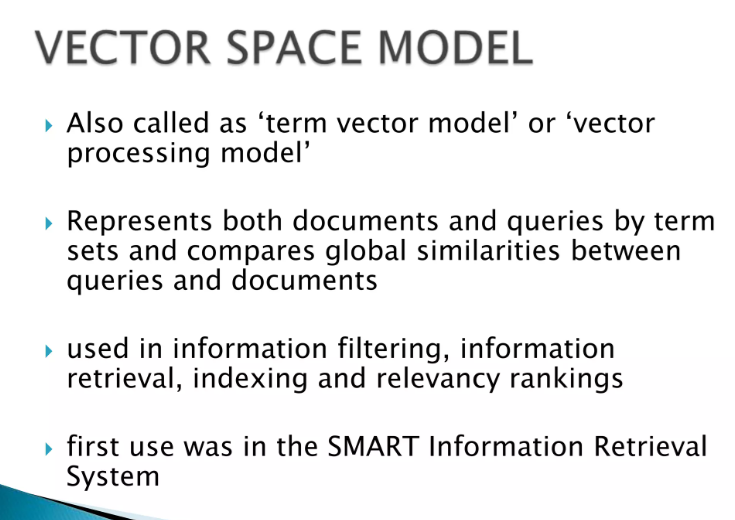


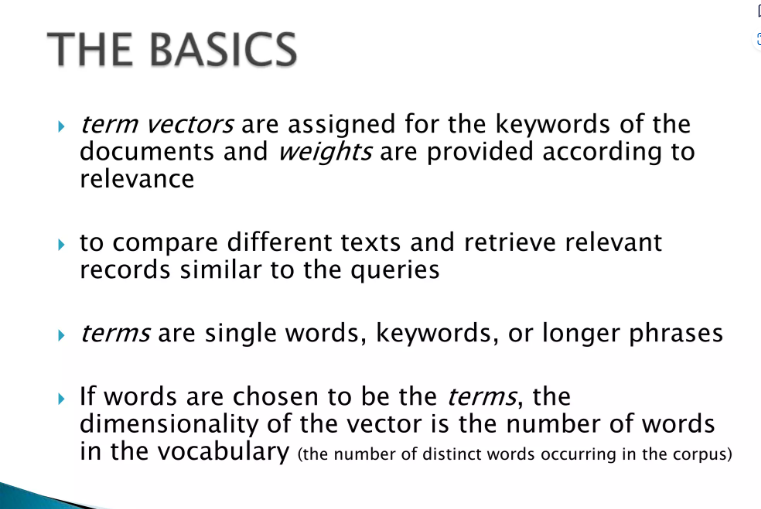
P(e) language model could be a trigram model, estimated from any data (parallel coupus not needed to estimate the parameters)

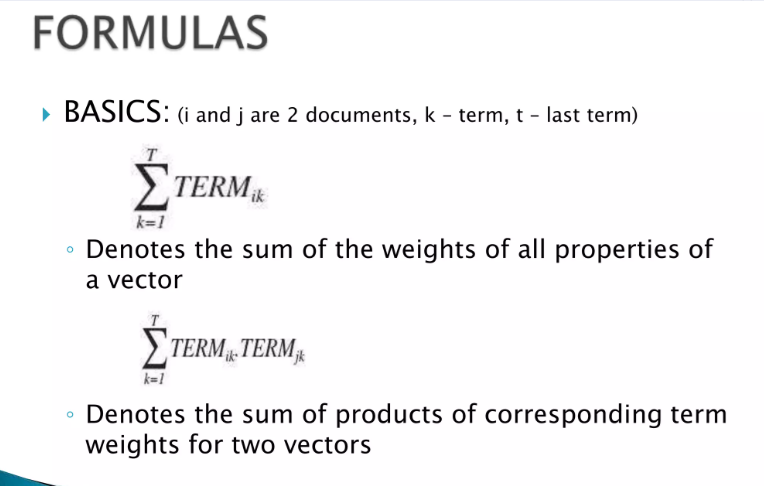
The translation model p(f|e) is trained from a parallel corpus of French/English pairs.

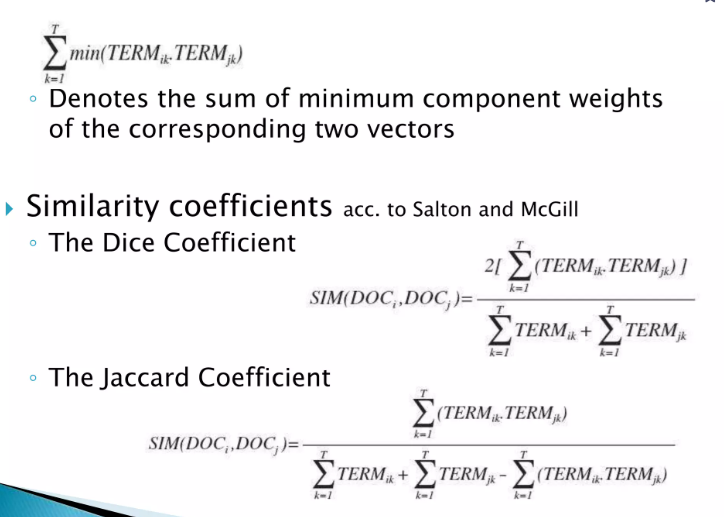


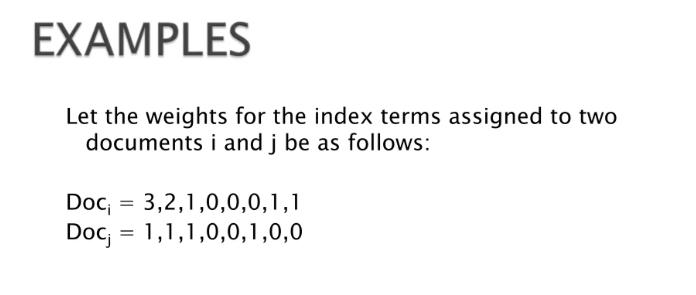


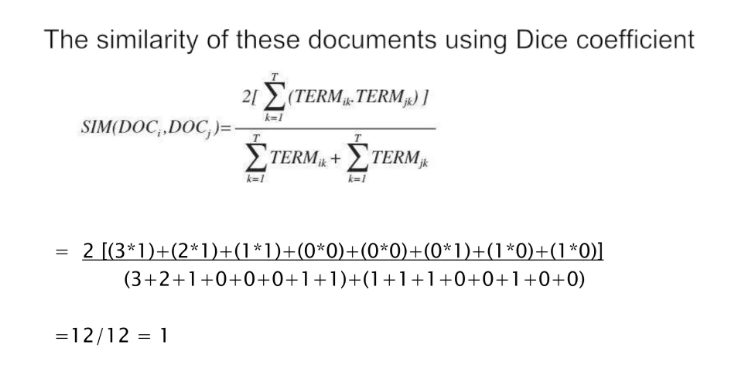
****

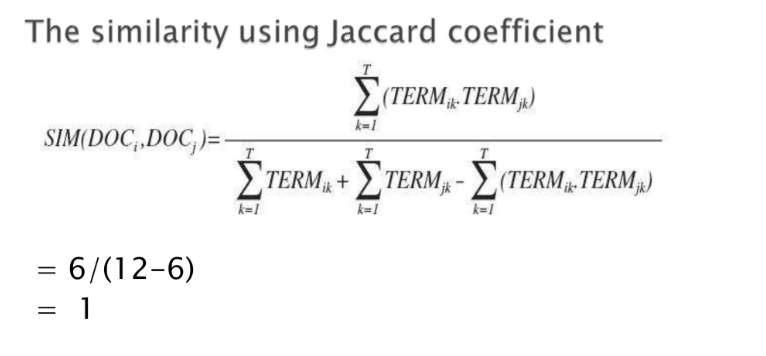
****

****

****

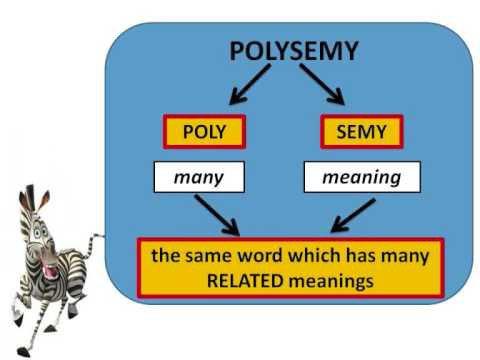
****

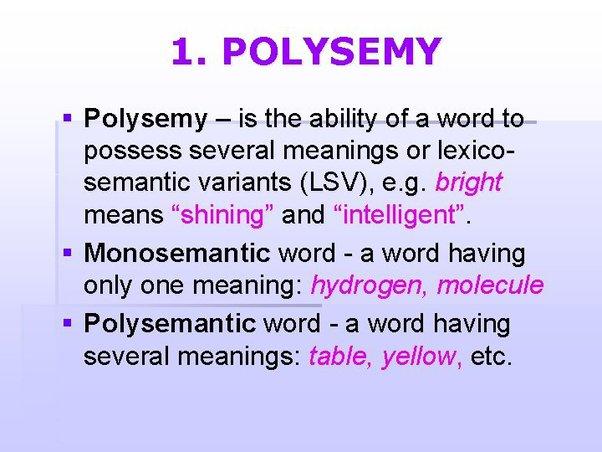
****

****

**homonymy, polysemy, synonymy**

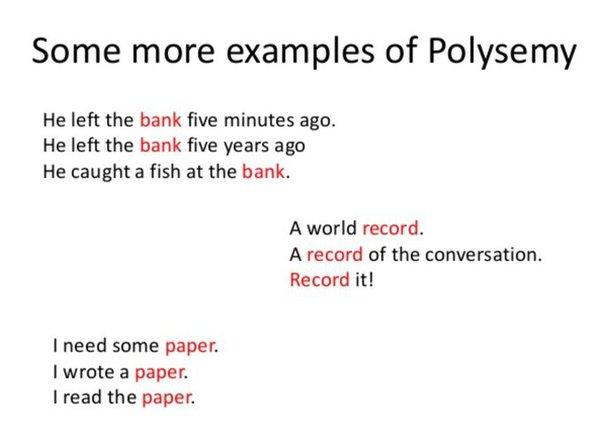
**Polysemy**is the association of one word with two or more distinct meanings, and a *polyseme*is a word or phrase with multiple meanings.





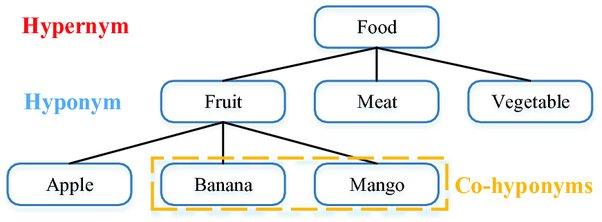
Here are a few examples of ***polysemous***words. They are shown first in a *primary-meaning* context and followed by a *secondary-meaning* context.

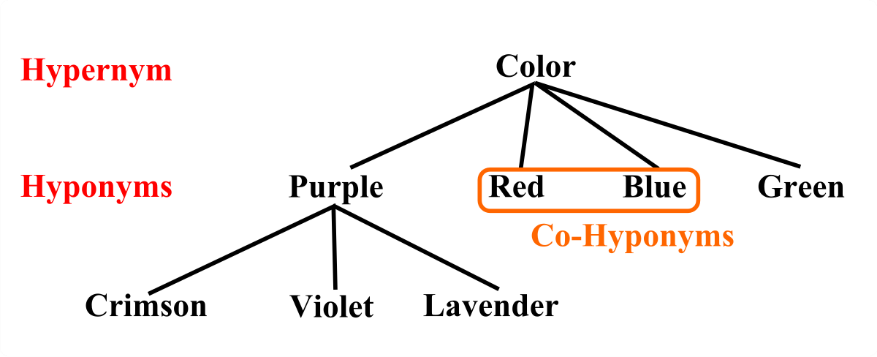
* *Arms* bend at the elbow.  
  Germany sells *arms* to Saudi Arabia.
* Boil the solution *once* with salt and *once* with sugar.  
  *Once* Germany had surrendered, the Soviets were free to enter the conflict against Japan.

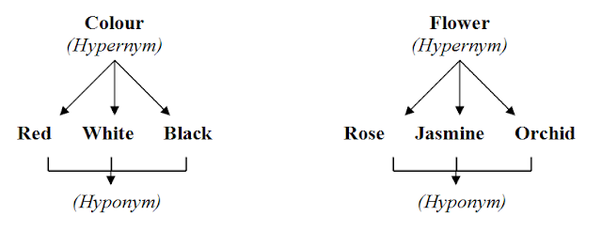


In linguistics and lexicography, ***hyponym***is a term used to designate a particular member of a broader class. For instance, *daisy* and *rose* are hyponyms of *flower*. Also called a *subtype*or a *subordinate term*.



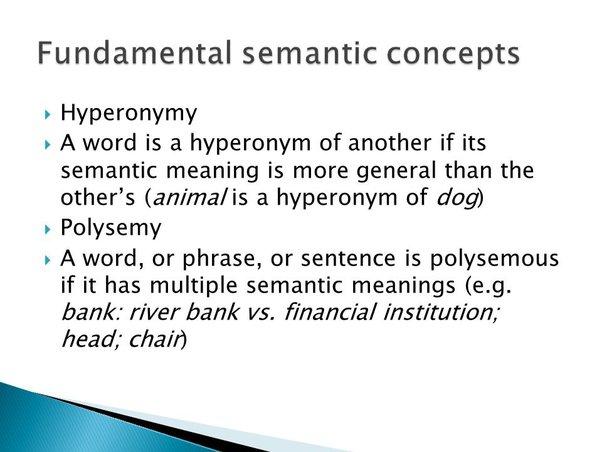


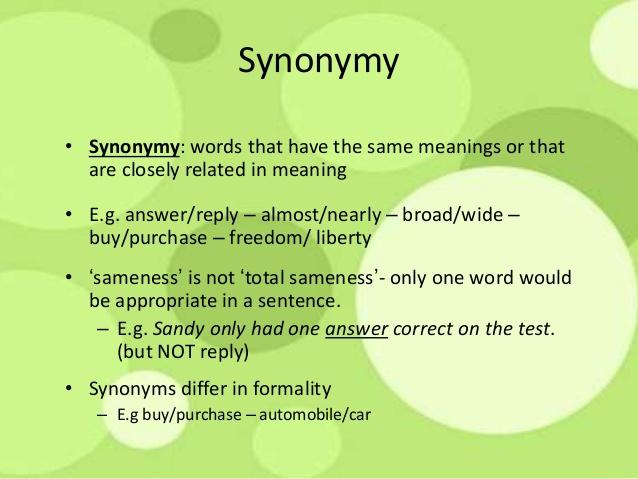


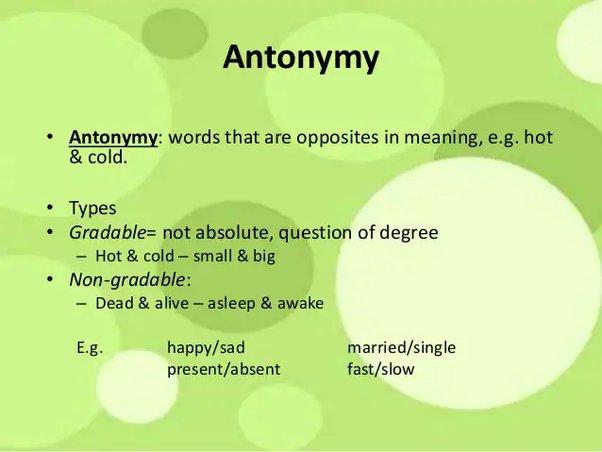


***Homonymy***- There are many pairs or groups of words, which, though different in meaning, are pronounced alike, or spelt alike, or both. Such words are called homonyms, and such words are in homonymy.

Lexical items which have the same sound or spelling or both, but differ in meaning, are called homonyms. Most of the homonyms are monosyllabic.







## User Query Improvement

The primary goal of any information retrieval system must be accuracy − to produce relevant documents as per the user’s requirement. However, the question that arises here is how can we improve the output by improving user’s query formation style. Certainly, the output of any IR system is dependent on the user’s query and a well-formatted query will produce more accurate results. The user can improve his/her query with the help of ***relevance feedback***, an important aspect of any IR model.

## Relevance Feedback

Relevance feedback takes the output that is initially returned from the given query. This initial output can be used to gather user information and to know whether that output is relevant to perform a new query or not. The feedbacks can be classified as follows −

### **Explicit Feedback**

It may be defined as the feedback that is obtained from the assessors of relevance. These assessors will also indicate the relevance of a document retrieved from the query. In order to improve query retrieval performance, the relevance feedback information needs to be interpolated with the original query.

Assessors or other users of the system may indicate the relevance explicitly by using the following relevance systems −

* **Binary relevance system** − This relevance feedback system indicates that a document is either relevant (1) or irrelevant (0) for a given query.
* **Graded relevance system** − The graded relevance feedback system indicates the relevance of a document, for a given query, on the basis of grading by using numbers, letters or descriptions. The description can be like “not relevant”, “somewhat relevant”, “very relevant” or “relevant”.

### **Implicit Feedback**

It is the feedback that is inferred from user behavior. The behavior includes the duration of time user spent viewing a document, which document is selected for viewing and which is not, page browsing and scrolling actions, etc. One of the best examples of implicit feedback is ***dwell time***, which is a measure of how much time a user spends viewing the page linked to in a search result.

### **Pseudo Feedback**

It is also called Blind feedback. It provides a method for automatic local analysis. The manual part of relevance feedback is automated with the help of Pseudo relevance feedback so that the user gets improved retrieval performance without an extended interaction. The main advantage of this feedback system is that it does not require assessors like in explicit relevance feedback system.

Consider the following steps to implement this feedback −

* **Step 1** − First, the result returned by initial query must be taken as relevant result. The range of relevant result must be in top 10-50 results.
* **Step 2** − Now, select the top 20-30 terms from the documents using for instance term frequency(tf)-inverse document frequency(idf) weight.
* **Step 3** − Add these terms to the query and match the returned documents. Then return the most relevant documents.