Morphological Tagging to Resolve Morphological Ambiguities

Birocheau Gaëlle
Centre Lucien TESNIERE, Université de FRANCHE-COMTE, BESANCON, FRANCE
10, rue Léonard de Vinci, 25000 BESANCON, FRANCE
gaelle.birocheau@univ-fcomte.fr

Abstract
The issue of this paper is to present the advantages of a morphological tagging of English in order to resolve morphological ambiguities. Such a way of tagging seems to be more efficient because it allows an intention description of morphological forms compared with the extensive collection of usual dictionaries. This method has already been experimented on French and has given promising results. It is very relevant since it allows both to bring hidden morphological rules to light which are very useful especially for foreign learners and take lexical creativity into account. Moreover, this morphological tagging was conceived in relation to the subsequent disambiguation which is mainly based on local grammars. The purpose is to create a morphological analyser being easily adaptable and modifiable and avoiding the usual errors of the ordinary morphological taggers linked to dictionaries.

Introduction
“The usefulness of corpora as a resource for language related research is proportional to the degree to which they have been linguistically annotated. It is much harder to retrieve interesting facts from a raw corpus than from one in which tokens have been marked for their wordclass, or one in which the syntactic structure of each utterance has been determined.”

1. Grammatical Categories
1.1. Generalities
NLP has to deal with large variations about grammatical categories. Dictionaries don’t often agree on the number and nature of them. That’s the same in NLP devices where categories also differ in number and nature depending on their purpose.

Looking at four dictionaries results tagging words – two of them are paper dictionaries, others are electronic ones – we can see the lack of norms through the great variations but could we talk about norms with natural languages? (see Table1, Annex1)

By observing the entries and categories of chosen words in the two former bilingual dictionaries we can see two major differences:
- first of all parts of speech are not always the same,
- and, when they are the same, they can appear in a different order.

We can guess that it comes from the internal priorities lexicologists gave within a particular dictionary dedicated to a special usage. For instance Robert & Collins seem to make the parts of speech (when several) appear in relation to the decreasing frequencies whereas Harraps gives all the possible tags in a predefined order (N, Vtr, Vi, Mod, V substitute, Vaux…). At least we can guess that the order is based on the relative frequencies of each category.

But, even if these methodologies seem both to be based on frequency, their results could be greatly different: there are much more nouns than modals in language but what about ambiguous forms? The word CAN, for example, is much often a modal than a noun.

1.1.1. Traditionnal Grammatical Partition
Concerning the first difference the problem is all the more difficult that linguists have never agreed on the number and nature of grammatical classes although grammatical categorisation has been studied since Antiquity.

Antiquity gave 8 traditional parts of speech but modern linguistics refuses this partition which is not based on a strong theoretical basis since the partition is based on a multiplicity of heterogeneous criteria – formal (morphological), notional (semantical, psychological or logical) and functional (syntactical) ones.

Some say that traditional partition has a real practical efficacy besides its unpleasant theoretical foundation. Others see it as a great intuition but can’t see any possible application.

Thus we can see a lot of adaptations ranging from 2 to 23 (or more) parts of speech coming from the 8 fundamentals.

1.1.2. Grammatical Partition In Nlp
NLP introduced new needs and new criteria and we assist at an increasing number of the grammatical POS (just look at the tags used by the electronic OED over there which not only gives the nature of words but also the function) in the devices. A large tagset is often used to prevent ambiguities and interferences in devices with a special purpose and it often runs. But it prevents from establishing pure theoretical rules, which could be used in many areas.

1.2. Practical Choice
In front of such a vagueness and since our aim wasn’t to set a theory on how to establish grammatical parts, we adopted the classification of the electronic OED with some regularisations. In one way it seems to be important for NLP tools to be based on existing supports because it could bring the possibility to adapt them on other supports and maybe on other languages – it brings adaptability to the system.

A large tagset could bring very good results on a very restricted field but what could that bring in a more general area? Using your own tagset can help you to avoid difficulties but on the other hand it implies building your own dictionary – the same tagset also brings functionality.

---

1 [10]
2 NLP: Natural Language Processing
3 [6]
According to me NLP cannot allow itself building a dictionary for each different purposes under the risk of losing efficiency. It must tend toward “a kind of normalisation” (even if it’s a utopian view in the field of natural language) and if not, try to use the existing tools to exploit them and have the same point of departure.

We finally adopted the traditional partition (8 categories) divided into subdivisions as the OED does.

2. The Data

We chose the Oxford English Dictionary because:
- Its quality is agreed.
- It includes large data.
- Its electronic form makes the treatment easier.

2.1. Data Analysis

2.1.1. Data Extraction

We studied the 8 traditional classes.

We first extracted the four files of the four major categories:
- Adjectives,
- Adverbs,
- Nouns,
- Verbs.

Then we joint the four minor categories:
- Conjunctions,
- Interjections,
- Prepositions,
- Pronouns.

Indeed, even if the relatively small number of the words contained in these categories could allow to set a complete dictionary with them we thought interesting to analyse their morphological behaviour to check whether their structure is less logical than in the four major classes - as it seems to be - or if the structure depends on a deep logic invisible from the surface.

The further results will show if we can set out some regularities in the structuring rules or if not if we should implement them totally.

Concerning the determiners we set them aside deliberately because of the own definition we wanted to give to them. In fact we can notice that even if there are no unilateral definitions of the classes, linguists relatively agreed on major classes (adjectives, adverbs, nouns and verbs).

But that’s not the same with minor classes that’s why we followed the OED description for the minor classes (conjunctions, interjection, prepositions and pronouns) but not for the determiners which will be redefined in relation to the further analysis. The class will be totally implemented further.

2.1.2. Data Pretreatment

We had to clean and check the files to eliminate the semantic information that could give birth to noise in an unsemantical analysis.

Some errors remains but they only concern 0.04% of the ambiguous forms and 0.03% of the whole data. So we consider that such error rates are negligible and couldn’t question the following results.

Here is the final repartition of the eight categories:

| P.O.S  | WORDS NUMBER |
|--------|--------------|
| adjectives | 48355        |

2.2. Data Processing

We then processed data under ACESS with requests

2.2.1. Establishing The Various Categories

The fist processing was the crossing of the eight classes to find the rates of the ambiguities in the corpus.

Crossing the eight simple classes made us set out 28 double classes coming from the addition of the combinations of the simple classes:

\[ C_8^1 + C_8^2 + C_8^3 + C_8^4 + C_8^5 + C_8^6 + C_8^7 + C_8^8 = 7+6+5+4+3+2+1 = 28 \]

The double classes have been crossed with the simple ones to give 56 triple classes:

The procedure was the same to obtain the quadruple (15), the quintuple (6) and the six fold (1) classes.

We didn’t find a word belonging to more than six categories at the same time.

2.2.2. Results For The Most Numerous Simple, Double And Triple Classes:

(In quantitatively decreasing order)

| SIMPLE CLASSES | WORDS NB | DOUBLE CLASSES | WORDS NB | TRIPLE CLASSES | WORDS NB |
|----------------|----------|----------------|----------|----------------|----------|
| Adj            | 48355    | Noun/Verb      | 6544     | Noun/Adj/Verb  | 592      |
| Verbs          | 24776    | Verb/Adj       | 1712     | Adv/Adj/Verb   | 172      |
| Nouns          | 15306    | Noun/Adj       | 1622     | Adv/Verb       | 85       |
| Adv            | 7826     | Adv/Adj        | 700      | Adv/Noun       | 61       |
| Interj         | 629      | Verb/Adv       | 233      | Noun/Verb/Prep | 44       |
| Prep           | 192      | Noun/Prep      | 178      | Noun/Verb/Prep | 23       |
| Pro            | 134      | Interj/Verb    | 87       | Interj/Adj/Verb| 18       |
| Conj           | 83       | Interj/Noun    | 83       | Noun/Adj/Interj| 14       |
| TOTAL          | 97301    |                | 11359    |                | 1107     |

Table2 :

Adj: Adjectives
Adv: Adverbs
Interj: Interjections
Prep: Prepositions
Pro: Pronouns
Conj: Conjunctions

We can see that the major classes are the more numerous but it was the way we defined them. The biggest class is the adjectives’ one.

As for the double classes four categories represent 93.12% of all the double categories: Noun/Verb: 57.61%

5 A simple class (ex: Adj) is an ambiguous class which contains words belonging at least to the part of speech maning the class (adjectives).

6 A double class (ex: Noun/Verb) is an ambiguous class which words belong at least to two parts of speech (Noun and Verb).
We can see that they only are combinations of the major classes. In the triple categories the four biggest ones constitute 82.2% of them:
- Noun/Adj/Verb: 53.48%
- Adv/Adj/Verb: 15.54%
- Noun/Adv/Verb: 7.68%
- Adv/Adj/Noun: 5.5%

Once again they only are combinations of the major classes what is quite logical mathematically: the most numerous double class crossed with the most numerous simple one should make the most important triple class.

2.3. Ambiguities Determination

Of course the first analysis is not sufficient to set out the ambiguities. It only allowed us to constitute ambiguous classes since they both contain ambiguous and non-ambiguous forms. For instance the Adjectives include the Nouns/Adjectives; Nouns/Adjectives/Verbs is included in the same time in Nouns/Verbs, Nouns/Adjectives, Verbs/Adjectives, Adjectives, Verbs and Nouns:

New requests allowed us to extract non-ambiguous classes.

2.3.1. Some Results

| Simple classes | Total | NB non ambiguous | % non ambiguous | Nb Ambiguities | % Ambiguities |
|----------------|-------|------------------|-----------------|---------------|--------------|
| Adj            | 48355 | 45101            | 93.27           | 3254          | 6.73         |
| Verbs          | 24776 | 17163            | 69.27           | 7613          | 30.73        |
| Nouns          | 15306 | 7641             | 49.92           | 7665          | 50.08        |
| Adv            | 7826  | 6985             | 89.25           | 841           | 10.75        |
| Interj         | 629   | 491              | 78.06           | 138           | 21.94        |
| Prep           | 192   | 140              | 72.92           | 52            | 27.08        |
| Pro            | 134   | 109              | 81.34           | 25            | 18.66        |
| Conj           | 83    | 72               | 86.75           | 11            | 13.25        |
| TOTAL          | 97301 | 77702            | 79.86           | 19599         | 20.14        |

Table 3

| Double Classes | Total | NB non ambig | % non ambig | Nb Ambig | % Ambig |
|----------------|-------|--------------|-------------|----------|---------|
| Nom_Verb       | 6544  | 5642         | 86.22       | 902      | 13.78   |
| Verb_Adj       | 1712  | 842          | 49.18       | 870      | 50.82   |
| Nom_Adj        | 1622  | 1374         | 84.71       | 248      | 15.29   |
| Adv_Adj        | 700   | 345          | 49.29       | 355      | 50.71   |
| Verb_Adv       | 233   | 72           | 30.9        | 161      | 69.1    |
| Nom_Adv        | 178   | 88           | 49.44       | 90       | 50.56   |
| Interj_Verb    | 87    | 29           | 33.33       | 58       | 66.67   |
| Interj_Nom     | 83    | 63           | 75.9        | 20       | 24.1    |
| Conj_Adj       | 5     | 1            | 20          | 4        | 80      |
| Conj_Adv       | 5     | 1            | 20          | 4        | 80      |
| TOTAL          | 11235 | 8463         | 75.33       | 2746     | 24.44   |

Table 4

2.3.2. Analysis

Table 3

The high ambiguity rates shows that the ambiguities in these classes are often included in other classes at the same time.

We also can see that the minor classes are often over ambiguous (Conj_Adj: 80%, Conj_Adv: 80%) even if not numerous.

Contrary to this are the simple classes Nouns and Verbs, which are very ambiguous but essentially form the double class Noun_Verb that is not very ambiguous (13.78%).

This double class, which contains 6544 members among which 5642 are non-ambiguous ones, includes the great majority of the ambiguous verbal (7613) and nominal (7665) forms.

2.4. Analysing Ambiguities

Now we analysed ambiguities relatively to their morphological forms so that we can set out grammatical predictive morphological rules.

3. Morphological Tagging

Then we could process data.

3.1. Morphological Rules

3.1.1. Simple Surface Criteria

We firstly established simple surface criteria.

The figure shows the distributions of the major classes under the first criteria. (see Figure 1)

Indeed, we could see that minor classes were never in majority whatever the criteria could be. Moreover such classes present high ambiguity rates. We therefore decided to implement them totally to prevent noise since it
seemed impossible to set out morphological rules to their description.

3.1.2. Rules Formal Definition
Establishing rules requires a formal definition. Here are the formal criteria used to set the rules:
a) We firstly consider the non-ambiguous forms only.
b) In any case: the number of exceptions ≤ the number of forms treated by the rules.
c) There’s a rule when the most important number of forms is at least ≈ (± 10%) to the double of the just inferior number. The most important number is then considered as the rule and the rest treated as exceptions.
d) When we are in front of small quantities (<10 in each category) we just need to check the second condition to establish a rule.
e) Without clear rules, small quantities are treated as exceptions.
f) In other cases, we must look for deeper criteria to define rules.

3.1.3. THE RULES
We finally found:

| Criteria | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|----------|---|---|---|---|---|---|-------|
| Rules    | 9 | 100 | 643 | 823 | 226 | 51 | 1852 |
| Exceptions | 0 | 90 | 465 | 762 | 280 | 90 | 1687 |
| Total    | 9 | 190 | 1108 | 1585 | 506 | 141 | 3539 |

Table 5
Be careful that the figures over there are not the number of forms but the number of rules and in one way exceptional rules. Of course, the number of exceptional rules is nearly the same as the number of rules but, from the formal definition of the rules, rules recover very more forms than exceptional rules. The 6 criteria allow recognising all the forms of the corpus.

3.2. Derivation Rules
We joint derivation rules to the morphological rules.

3.2.1. Derivation Rules Applied To Verbal Forms
- Past participles
- -ING forms
- Tenses
- Present tense
- Contracted forms

3.2.2. Derivation Rules Applied To Nominal Forms
- Plural
- -ING forms
- Collective noun
- Genitive forms

3.2.3. -ING forms
Besides -ING forms included in progressive verbal tenses are gerundive forms, which tend to loose their verbal nature to take a nominal one as a gerund (I like swimming) or an adjective (Sleeping Beauty).

4. Conclusion
4.1. The Morphological Tagging
⇒ The morphological tagging has already been implemented on French in the laboratory.
⇒ The morphological tagging of English gives promising results since it allows to take the neology into account. Descriptive rules could also be used by foreign learners to have a deep view of the morphological structure of the foreign language. That’s why we can plead in favour of our descriptive linguistic model contrary to statistical or probability models.
⇒ Another advantage is the system of rules which prevents from having a complete dictionary. That’s very useful as regards computing since speed is a required condition to have an effective system especially on the web.
⇒ The research was of course first limited to simple words but could easily be extended to compounds. That is what is now being done on French in our laboratory.

4.2. Methodology
The methodology presents several advantages.
⇒ First the method uses existing supports what prevents us from rebuilding a complete dictionary.

- Gerund defined as a nominalization of a predicate.
- Verbal Nouns defined as a nominalization of a lexical verb.
But problems remain with sentences like:
I hate his telling lies to you / I hate him telling lies to you.
So the theoretical problem about -ING nominalizations is not resolved yet.
We thus chose to have a very basic definition in disambiguating non-verbal -ING forms:
I hate his telling lies to you.
Telling will be considered to have a nominal nature (to be a noun) since it follows a possessive adjective (determiner).
I hate him telling lies to you.
Telling will be considered as having a verbal nature (being a verb) since following a personal pronoun.
In fact, as we saw in the first part, grammatical categorization is not theoretically resolved. So we can consider as Jean TOURNIER does:
⇒ Le lexicographe ayant un dictionnaire à faire est tenu de résoudre par lui-même, comme il le peut, souvent de façon empirique et subjective, des problèmes que la linguistique théorique n’a pas encore résolus. C’est notamment le cas du problème des classes de mots, neuf selon les uns, huit selon les autres, deux selon Vendryes, cinq selon une plus récente analyse. \[p172\]

4. [1]
10 Remember we don’t care about semantics!
11 [7]
12 The lexicologist who has a dictionary to do must resolve by himself the problems remaining inside theorical linguistics and specifically the problem of P.O.S.
⇒ Second it can be reproduced (it has been applied to French and to English) what offers the possibility to apply it on other Indo-European languages.
⇒ Finally the steps have been described very precisely what could permit an automation of the method on a chosen language at a chosen level from a chosen support.

4.3. The Disambiguation System
But, even if the models based on n-grams probabilities like CLAWS 2 claim to have 96-97% success rates, if statistical taggers based on HMM usually reach 96%, we recently saw a disambiguiser based on constraint rules with a precision of 99.7% what reinforced our belief in linguistic models.

4.3.1. Local Grammars
Till now, “no formal theory is able to take the whole syntactic complexity into account.” Moreover, it seems that “Most of the morphological ambiguities could be disambiguated by only having a close look at their local context.” Thus we chose to build local grammars to resolve morphological ambiguities.

We used the BNC annotated by CLAWS to extract the various contexts of the ambiguities. Then we wrote local grammars able to assign the right tag to an ambiguity in a specified context.

4.3.2. The British National Corpus
Consulting the BNC annotated by CLAWS makes us find a lot of errors.

**RECOGNITION ERRORS**

| LEXEMES | VVB | VVI | VM0 | NN1 | Nb of found sol | Real Nb of sol | Diff |
|---------|-----|-----|-----|-----|-----------------|----------------|------|
| work    | 5007| 19297| 57402| 81706| 91355           | 9649           |      |
| take    | 19694| 51486| 153| 71333| 71735           | 402            |      |
| surprise | 34  | 465  | 4673 | 5172 | 5357            | 165            |      |
| arrive  | 1100| 1802 | 0  | 2902 | 2912            | 10             |      |
| program | 15 | 68 | 3873 | 3956 | 4062            | 106            |      |
| programme | 9  | 10  | 18936 | 18955 | 19071           | 116            |      |
| can     | 734 | 9    | 234386 | 1019 | 236148          | 7173           |      |
| will    | 1228| 24   | 244823 | 6392 | 252467          | 2100           |      |
| want    | 31871| 24858| 0  | 421 | 57150           | 57547          | 397  |

**Table 6:**

VVB = The finite base form of lexical verbs
VVI = The infinitive form of lexical verbs
VM0 = Modal auxiliary verb
NN1 = Singular common noun

We can notice data loss rates ranging from 0.34% – negligible- to 10.36% -which is not inconsiderable! -. Recognition errors can either come from the initial tagging or the disambiguation (re-tagging).

**DISAMBIGUATION ERRORS**

Here are examples of sentences given by the BNC when I looked for:

1. works as being a VVB - conjugate verb -
2. works as being a VVZ- the third singular person of the present tense –

| LEXEMES | VVZ | NN1 | Nb of found sol | Real Nb of sol | Diff |
|---------|-----|-----|-----------------|----------------|------|
| works   | 6157| 30  | 6187            | 14528          | 8341 |
| takes   | 11674| 16  | 11690           | 11823          | 133  |
| surprises | 57 | 365 | 422            | 461            | 39   |

**Table 7:**

NN1 = Singular common noun
VVZ = The -s form of lexical verbs

We can notice data loss rates ranging from 0.11% – negligible- to 51.41% -which is not inconsiderable! -. Recognition errors can either come from the initial tagging or the disambiguation (re-tagging).

**Table 8:**

VVB = The finite base form of lexical verbs
VVZ = The -s form of lexical verbs

We also encountered errors obviously coming from the disambiguation.

EX N° | LOOKING FOR AS | SENTENCES PROPOSED BY THE BNC
1 | work | VVB | ...studies of Gauguin’s work, ... |
2 | -- painters whose work we’re familiar with. |
3 | ...which supports whatever work we do in the organization. |
4 | works | VVZ | It has large brick works, engineering works and freezing factories. |
5 | ...that Jenkins composed these works in his 20s or 30s.... |
6 | Are all the artists showing works specially created for ‘Documenta’? |

13 Hidden Markov Model
14 [14]
15 [2]
16 [9]
We briefly analysed the erroneous examples, tried to give a basic explanation of the problem and suggested a solution:

| EX N° | PROBLEM | POSSIBLE SOLUTION |
|-------|---------|-------------------|
| 1     | genitive | Npr’+s+N/V ⇒ N   |
| 2     | Possessive phrase | N+whose+N/V ⇒ N |
| 3     | Agreement in number | Whatever+N/V ⇒ N |
| 4     | Plural + coordination | factories=N ⇒ works=N |
| 5     | ? | N+V+D+N/V+Prep ⇒ N |
| 6     | -ING form | if ING=V, N+V+N/V+Adv+Adj ⇒ N |

Table 9:

Npr = Proper Noun
N/V = Noun/Verb ambiguity
D = Determiner
Prep = Preposition
Adj = Adjective
Adv = Adverb

The local grammars method seems to be able to give good results on these special problems.

5. Conclusion

There is an increasing need for linguistically annotated corpora but a lack of such available corpora. We now achieve good precision with statistical models but we think linguistics can’t be satisfied with this solution. So we chose to create and implement a linguistic morphological model to disambiguate morphological ambiguities.

The morphological tagging seems to be more efficient because it gives an intention description of morphological forms compared with the extensive collection of usual dictionaries.

It is very relevant since it allows both to bring hidden morphological rules to light which are very useful especially for foreign learners and take lexical creativity into account.

The very precise description of the methodology can drive to an automation of the procedure.

The morphological disambiguation is based on local grammars since there is no complete definite linguistic theory able to describe all the syntax complexity.

Moreover local grammars offer more flexibility than a global syntactic model.

Of course we first didn’t take compounds into account, we are certainly going to let some syntactical structures aside and the corpus is obviously limited but the model seems to be linguistically efficient.

Our purpose is to create an easily adaptable and modifiable morphological analyser which can avoid the usual errors of the ordinary morphological taggers and disambiguate the most refractory ambiguities.

It could be used either as the first step of a more complete analyser, either under its descriptive form, or to annotate corpora.

6. References

[1] COTTE P. (1994). Le paradoxe du nom verbal en anglais contemporain. Université Charles de Gaulle - Lille III, SILEX URA D382 du CNRS. In BASSET L. & PERENNÉC M. Les classes de mots, tradition et perspectives (pp 233-264). Lyon, PU de Lyon.
[2] FUCHS C. (1988). L’ambiguïté et la paraphrase : opérations linguistiques, processus cognitifs, traitements informatisés. Université de Caen, Catherine Fuchs (Ed.).
[3] FUCHS C.(1993). Linguistique et traitement automatique des langues, Baume-les-Dames, Hachette.
[4] GROSS G. (1988). A quoi sert la notion de partie de discours ? In Langages N°92 (Décembre 1988). Les parties du discours. (pp 37-49).
[5] LE GUERN M. (1988). Parties du discours et catégories morphologiques en analyse automatique. In Langages N°92 (Décembre 1988). Les parties du discours. (pp 37-49).
[6] LAGARDE J.P. (1988). Les parties du discours en linguistique contemporaine In Langages N°92 (Décembre 1988). Les parties du discours. (pp 37-49).
[7] TOURNIER J. (1985). Introduction descriptive à la lexicogénétique de l’anglais contemporain. Champion-Slatkine, Paris-Genève.
[8] TOURNIER J. (1991). Structures lexicales de l’anglais. Poitiers, Nathan.
[9] Centre National de la Recherche Scientifique, Université Joseph Fourier-GrenobleI.(1987). B. VAUQUOIS et la TAO, 25 ans de Traduction Automatique. Grenoble, Ch. Boitet éditions.
[10] http://lands.let.hun.nl/projects/exploitation.en.html Quiering the BNC on line:
[11] http://thetis.bl.uk/cgi-bin/saraWeb?qy... Grammatical tagging of the BNC:
[12] http://info.ox.ac.uk/bnc/what/gramtag.html CLAWS to annotate the BNC:
[13] http://info.ox.ac.uk/bnc/what/garside_allc.html
[14] http://www.cse.ogi.edu/CSLU/HTLsurvey/ch3node3.html#SECTION3


## ANNEX 1

### Four Dictionaries Tagging Ambiguous Words

|                  | HARRAPS COMPACT | ROBERT & COLLINS | OED\(^\text{17}\) | BNC\(^\text{18}\) |
|------------------|----------------|-----------------|----------------|----------------|
| **Taking**       | Adj, N         | Adj, N          | Vbln, Ppla     | Vvg, Nn1, Aj0  |
| **Programming**  | N              | N               | Vbln           | Vvg, Nn1       |
| **Working**      | Adj, workings = N plu | Adj, workings = N plu | Vbln, Ppla     | Vvg, Nn1, Aj0  |
| **Doing**        | N              | N               | Vbln,         | Vvg, Nn1       |
| **Arriving**     |                |                 | Vbln           | Vvg, Nn1, Aj0  |
| **Surprising**   | Adj            | Adj             | Vbln1, Vbln2, Ppla | Vvg, Nn1, Aj0 |
| **Being**        | N              | N               | Vbln, Ppla     | Vbg, Nn1       |
| **Willing**      | Adj            | Adj, N          | Vbln, Ppla     | Vvg, Aj0, Nn1  |
| **Wanting**      | Adj, Prep      | Adj, Prep       | Vbln, Ppla, Pple | Vvg, Nn1      |
| **Take**         | N, Vtr, Vi     | N, Vtr, Vi      | N, V           | Vvi, Vvb, Nn1  |
| **Work**         | N, Vtr, Vi     | N, Vi, tr       | N, V           | Nn1, Vvi Vvb   |
| **Surprise**     | N, Vtr         | N, Adj, Vtr     | N, V           | Nn1, Vvi Vvb   |
| **Arrive**       | Vi             | Vi              | N, V           | Vvi, Vvb       |
| **Program**      | N (US), Vtr    | N (US), Vi, Vtr | Nn1, Vvi Vvb   |
| **Programme**    | N, Vtr         | N, Vtr          | N, V           | Nn1, Vvi Vvb   |
| **Can**          | N, Vtr, Mod    | Mod, N, Vtr     | N (2), V (3)   | Vm0, Nn1, Vvb, Vvi |
| **Will**         | N, Vtr, Mod    | Mod, Vtr, N     | N (4), A, V (3), Adv | Vm0, Nn1, ,Vvb, Vvi |
| **Want**         | Vi, Vtr, N     | N, Vtr, Vi      | N (2), V      | Vvi, Vvb, Nn1  |

**TABLE1:**

| Adj = Adjective | Prep = Preposition |
|-----------------|---------------------|
| Plu = plural    | Vtr = transitive verb |
| V = Verb        | Mod = Modal        |
| Vi = intransitive verb | Ppla = participle used as an adjective |
| Vbln = -ING form | Pple = participle used in verbal form |
| Adv = Adverb    | Nn1 = singular common noun |
| N = Noun        | Vvi = The infinitive form of lexical verbs |
| Vvb = The finite base form of lexical verbs | Vm0 = Modal auxiliary verb |
| Vdg = The –ing form of the verb DO: Doing | Vbg = The –ing form of the verb BE: Being |
| Vvg = The –ing form of lexical verbs | Aj0 = Adjective |

---

\(^{17}\) OED: Oxford English Dictionary  
\(^{18}\) BNC: British National Corpus (tagged by CLAWS)
