# Protocol for Orthographic Transcription

(Revised: September 27th 2000)

## A1. General introduction

The *Corpus Gesproken Nederlands* (CGN) is a joint Flemish-Dutch project with the aim of collecting around 10 million words of spoken Dutch. The project is financed by the Flemish and Dutch governments and the Dutch Organization for Scientific Research, (NWO).

Technologically, such a corpus is essential for the further development of Dutch language and speech technology, and consequently helps secure the future of Dutch as a cultural language in a muti-lingual Europe. CGN is also of particular importance for other fields of research: for example lexicography, language teaching, speech and language development in children, sociolinguistics, psycholinguistics, phonetics, phonology and conversation analysis.

To make the corpus suitable for use in different fields of research, the fragments of speech need to be given an orthographic transcription. This protocol covers the rules governing such an orthographic transcription.

## A2. Definition of the orthographic transcription

An orthographic transcription is a *literal* reflection of what was said. The transcription corresponds closely to *written* Dutch. However, on certain points the *Dutch spelling rules* are deviated from or additional information is added. This ensures that the transcription is as consistent as possible and applicable for a range of scientific purposes.

The orthographic transcription also contains *alignment*[1] at *chunk level*[2] and an indication of the different speakers. The transcription also allows for the registration of background sounds and there is a field for comments.

[1] *alignment:* linking a part of the transcription to a part of a sound signal by inserting boundaries. These boundaries determine the start and end times of the parts of the transcription. See the manual supplied with the Praat program for further information over inserting boundaries.

[2] *a chunk:* a fragment of speech of approximately 2 to 3 seconds which is bordered on both sides by a (short) visible and audible pause. Chunks do not need to correspond to sentences or parts of sentences; they are determined solely by the pauses in the speech signal. You are encouraged to keep the chunks as short as possible, (where the sound signal permits). Further guidelines are given in the manual for the Praat program.

# B. Principal rules

B1. Use the new spelling. Apply the new rules also to the placing of hyphens (e.g.: zee-engte), the inserted-n (e.g..: panne**n**koek), etc. (see also F2).

B2. Work as accurately as possible, but **don't waste time** on sections which are difficult to understand or where people are talking over one another; mark these with **xxx** or **Xxx** (see C9 and C10).

B3. **Do not use a capital letter** at the beginning of a sentence. **Do use capital letters** for proper names **(**e.g.: Bert Jansens) and for titles of books, records, etc. (e.g.: De Naam Van De Roos) (see also C1 and C2).

B4. Use the following codes where necessary or logical:

- use ***v** for **foreign words** (e.g.: tomorrow*v) (see also C3);
- use ***d** for **dialect words** (e.g.: kortewagen*d; keuje*d) (see also D4);
- use ***z** for **standard words** which are pronounced with a **strong dialect** (see D5);
- use ***n** for **new words** (e.g.: zerotolerantie*n) (see also C4);
- use ***t** for **new interjections** (e.g.: amaai*t; hoppa*t) (see also C5);
- use ***a** for **words cut short or interrupted** (e.g.: uitges*a; verpr*a) (see also C6);
- use ***u** to show **onomatopeia** and **slips of the tongue** (e.g.: boink*u; gespreken*u) (see also C7);
- use ***x** for words where you are **not** sure if you have **understood** them **correctly** (see C8).

Never use more than one of the above codes per word. Choose the most suitable code.

B5. Check if the transcription only includes words which are in the CGN lexicon[3]. This can be done using the special CGN spelling checker.

B6. *Only* use the punctuation marks **full stop**, **question mark** and **ellipsis** ( **.** / **?** / **...** ). Do not use any other punctuation marks such as comma, exclamation mark, quotes etc. (see also section E).

**If in doubt:** Where the protocol is insufficient, choose the spelling which best fits the conventional spelling.

B7. Use **ggg** for **clearly audible *speaker's* sounds** (see also C11).

B8. Use the ***speaker_unknown*** tier[4] for parts where you don't know (for sure) which of the speakers has said them and for speech from a speaker who only says a few words and for whom there is no need to create a separate tier.

B9. Use the ***background*** tier for localizable **background sounds** which can be clearly heard and/or affect the course of the discussion (see C12).

B10. Use the ***comment*** field for all kinds of comments about the recording and for comments regarding background sounds which can be heard throughout the entire recording (see C13).

[3] *CGN lexicon:* a word list with all the words which we regard as "Standard Dutch". The list is however incomplete and will be added to during the course of the project (see "*n").

[4] *a tier:* a tier looks like a music stave or a line on which all the speech of a particular speaker is transcribed, (including the alignment at chunk level (see A2)). Every speaker has such a tier. Together all these tiers or (musical) staves visualize the course of the discussion and the interactions. This is comparable with a music score or theatrical script (see also the manual for the Praat program)

# C. Further explanation of the conventions used

C1. **Use capital letters for proper nouns.** Proper nouns include: place names, personal names, brand names, company names, club names, names of committees, …

- ik woon in **d**e **D**orpsstraat in **D**en **H**aag.
- een **V**laamse schrijver; **N**ederlandstalig
- **P**orsche; **M**ercedes; **F**ord; **C**itroën
- **T**ransport **V**ereecken; het **t**ransportbedrijf **V**ereecken
- **F**okker **S**ervices; de **s**erviceafdeling **v**an Fokker
- **C**lub **B**rugge; de **v**oetbalclub **A**jax

Where a proper name consists of several words, write every word with a capital letter, even if this is not in line with the Dutch spelling rules[5].

- Romain **D**e Coninck; professor **V**an Veen
- Karel **D**e Kale; Karel **D**e Vijfde; Jan **Z**onder Vrees
- **R**oel **I**n **'t V**eld (evt.: **R**oel-in-'t-Veld);

Where **'s-**, **d'**, **l'**, ... (a lower case letter with an apostrophe) occur in proper names, write these with a lower case letter.

- **'s-G**ravenhage; mevrouw **d'A**ncona; de regel van **l'H**opital

Be consistent in your notation: once you have used a particular spelling, use the same spelling whenever the proper name reoccurs in the same fragment.

C2. **Use capital letters for titles of books, records, etc.** Where a title consists of several words, write all the words with a capital letter.

- ik heb Hugo Claus' boek Het **V**erdriet **V**an **B**elgië gekocht.
- Heb jij **D**e **N**aam **V**an **D**e **R**oos al gelezen?

C3. **Use *v for words from a foreign language:** i.e. words which (with that particular meaning or pronunciation) do not occur in the vocabulary of Standard Dutch but do occur in a foreign language.

- **he's*v a*v rebel*v** is Engels voor hij is een rebel.
- vertrek nu maar snel. **see*v you*v tomorrow*v**.
- **cave*v canem*v** is Latijn voor pas op voor de hond.

Foreign words in titles of books, films etc. are also marked with a *v.

- de band **New*v Kids*v On*v The*v Block*v.**
- heb je **La*v Vie*v En*v Rose*v** al eens gehoord?

Proper nouns from a foreign language are **not marked** with a ***v**.

- het liedje **Josephine** van **Chris Rea**.
- heb je **Grace Jones'** laatste hit al gehoord?

Foreign words which have become established in Dutch are also **not marked** with a ***v**.

- bij het opstarten van de **computer** krijg je een **jingle** te horen. **save** jij die **file** eens?
- zal ze **überhaupt** wel te weten komen waar **Freetown** ligt?

[5] With this spelling we are expressly deviating from the conventional spelling; this is to ensure the most consistent transcription and that proper names can be easily recognized as a single entity.

C4. **Use *n for new words**: these are words which you think are correct Dutch but, on the basis of the spelling checker, appear not (yet) to be included in the lexicon. This includes neologisms which are becoming gradually established in our language.

- **zerotolerantie*n, topprioriteit*n, millenniumprobleem*n, vijfprocentseis*n.**

C5. **Use *t for interjections.** Only those interjections which are not (yet) included in the CGN lexicon should be marked with a ***t**.

- **tsjakka*t** die zit.
- **amaaikes*t** wat gaat er nu gebeuren.

The CGN lexicon already includes a lot of interjections; these should **not be marked** with a ***t**. First try writing an interjection without ***t**. Only add the ***t** when the spelling checker objects, (or accept the suggestion given by the spelling checker).

- dat denk ik niet **hoor**.
- **sjonge** wat hebben we nou?
- **allee** dat meen je toch niet **zeker**.
- **awel** 't is goed **zulle**.
- **hé** kom jij 'ns even hier alsjeblieft.
- ik **uh** weet het niet goed **zuh**.
- **mmm dat weet ik nog zo niet.**

In order to make transcribing interjections **consistent**, here is a list of frequently occurring, short interjections. **Only** write them as they appear on the list.

| | | | | |
|---|---|---|---|---|
| ah | bwa | hu | oesje | tut (and also "tut |
| aha | eih, eikes | hum (and also "hum hum") | oh, o | uh, uhm |
| ai | goh | ee (and also "o jee") | oho | uhu |
| au | ha, haha | mm-hu | poeh | wauw |
| bah | hé, hè, hei | mmm | pst | woh |
| boe | ho | oeh, oei | sjt, sst | zuh, zulle |

If a word is interrupted by an interjection, (generally by "uh"), use hyphens to link the composite parts and the interjection to each other and add ***u** (see also C7c):

- heb jij het geld **over-uh-gemaakt*u?**
- ze heeft haar ring **ver-uh-uh-patst*u**.

C6. **Use *a for words cut short or interrupted.** Use this code when a spoken word, for what ever reason, is unfinished. This also applies to words which are cut short and then repeated in full, as in the third example below. Use *a also for mispronunciations which are cut off and not resumed, (compare C7c). Where only one sound is spoken, as in <m*a ik weet het niet>, treat the sound as a cut off word and use *a, even if you do not know *which* word is cut short. Where a word is repeated in its entirety, use **neither** a code **nor** a hyphen. E.g.<ik ik ik weet het niet>.

- ik ga **mo*a** overmorgen naar de tandarts.
- neem de **link*a** uh de rechter deur.
- ik moet **mo*a** morgen naar de tandarts.
- ik moet **overm*a**... ik moet overmorgen naar Tilburg.
- **m*a**… ik weet het niet.
- ze heeft **bel*a** boulimia nervosa.
- het goede weer zal een **voelpr*a voorproefje** zijn.

C7. **Use *u to show pronunciation.** Use this code when someone, for whatever reason, says a word which is not in the CGN lexicon, (with that meaning), and which does not fall under *one* of the other categories (*v, *d, *n, *t, *z, *x, *a). Typical examples are:

C7a. onomatopoeia and (deliberate) distortion:

- **boink*u** hoorde ik. hij was tegen de muur gevlogen.
- ik **spreuk*u** een beetje **Zweuds*u**.
- hij zegt voor de grap **toekenbas***u in plaats van boekentas.

C7b. slips of the tongue:

- k heb me **verspraakt*u**. dit was een **versprekeling*u**.
- mmm ik lust wel een **broekje*u** kaas.
- KPN annuleert fusie met Spaans bedrijf **alduns*u aldus** een kop in de krant.

C7c. all cases of slips of the tongue and resumptions **within a word**. In these cases, use *u in combination with hyphens to link the different parts of the words to each other. Words which are interrupted by an interjection or a speaker's sound (e.g. laughter) also fall into this category.

- hij is nog niet naar de kapper **gewee-weest*u**
- de verzetsstrijders van het **Kosovo-bevrijdingslever-leger*u**.
- de **jeugdlied-liefdes*u** van de schrijver.
- dat is de **mogelijkhee-heid*u**.
- heb jij dat spul nog over het gras **uitges-ge-uh-gesproeid*u**?
- dat is een **ongeluks-uh-baby*u**.
- ze heeft haar ring **ver-uh-uh-patst*u**.
- ze heeft haar ring **ver-ggg-patst*u**.

**Note:** Use ***u** when a speaker makes a slip of the tongue but ultimately does pronounce all the parts of the word in question. When the speaker interrupts a word and does not resume it in full or in part, i.e. where words are not pronounced in full, use ***a** (see C6).

C8. **Use *x for words where you are not sure if you have understood them correctly.** Use this code when, in spite of difficult intelligibility, you are almost sure what was said.

- hij kan **dertig*x** bladzijden per uur uit het hoofd leren.
- ze zei dat **Jan-Peter*x** nog zou komen.

C9. **Use xxx, -xxx, xxx- or -xxx- for incomprehensible utterances.** Use this code where there is a word, part of a word or a series of words which you have not properly understood and are unable to make anything of them.

- ik heb al **xxx** bladzijden gelezen.
- ik heb al **xxx-**enzeventig bladzijden gelezen.
- ik heb al achten**-xxx-**tig bladzijden gelezen.

You should also transcribe utterances which do not really belong to the actual conversation, which are difficult to transcribe or include lots of background noise with **xxx**.

C10. **Use Xxx with incomprehensible personal names or titles**. Use this code when you have not properly understood a word and are unable to make anything of it, but definitely know that a personal name or title was said.

- dit truitje heb ik bij **Xxx** gekocht.
- ze zei dat **Xxx** nog zou komen.

C11. **Use ggg for clearly audible *speaker's* sounds** such as laughing, crying**,** screaming, coughing, etc. Some examples:

- even m'n keel schrapen hoor. **ggg**. zo dat is beter.
- ken je die mop van die man die naar Parijs ging? nou hij ging niet. **ggg**.
  (i.e. do not transcribe "haha" or anything similar when someone laughs).
- spreker A: **ggg**. (do not transcribe "hatsjie" or anything similar if someone sneezes).
  spreker B: gezondheid.

**Note**: It is expressly **not** the idea to identify all sorts of **weak speaker's sounds**, such as:

- breathing in and out, a small cough, smacking the lips, clicking the tongue.
- lightly clearing the throat, quiet breaking sound in the throat at the start of an utterance,...

C12. **Use the *background* tier for clearly audible background sounds** where you think it could be useful to know that they are there and/or sounds which clearly influence the course of the conversation.

You certainly do **not** need to go to any trouble of trying to mark the beginning and end of these sounds exactly. Just make sure the borders are **approximately** correct.

A series of separate sounds which belong together (as content) should be marked as one, with one beginning and one end.

- You hear applause during or after a speech.
  → Make a chunk with "**applause**", the exact position of the beginning and end are unimportant.
- You clearly hear footsteps and then a key being inserted in the lock. One of the speakers says "ha, mijn vrouw komt thuis.".
  → Transcribe "**footsteps + key in lock**" as a single phrase, the exact times are unimportant.
- During a conversation of half an hour five minutes of drilling can be heard. The sound is interrupted every now an then for a few seconds.
  → Make just one chunk of about five minutes with the comment "**drilling**".
- You hear a dog barking loudly, this barking is clearly present during the conversation.
  → Transcribe this barking as a single item with "**dog barking**", even if the dog is quiet on and off.

**Note**: You should definitely **not** go to any trouble to mark **weak and uninformative sounds**, like **small items in the recording**, such as:

- a ballpoint pen clicking during a meeting
- something being dropped, like a ballpoint pen
- a chair creaking
- a clock ticking
- a squeak or creak in the recording

C13. Use the ***comment*** field for sounds which characterize the entire recording or for things you notice as you listen to it. The comment field should not include any information on time, (no beginning and end are marked). Here are some examples:

- background music can be heard throughout the entire recording
- every now and then mumbling can be heard from other speakers, spread over the whole recording
- the entire interview is recorded in a moving car
- in the second part of the recording a parrot can be regularly heard

# D. How should spoken language be written down?

D1. **Write down what is said literally.** This means that you should change nothing of what is said; this also applies to where the speaker uses **incorrect sentence constructions** or makes **grammatical mistakes**:

| *you hear and write* | *do **NOT** correct to* |
|---|---|
| de koe **dat** graast.<br>**de** open venster.<br>**hun** zijn er al. | de koe **die** graast.<br>**het** open venster.<br>**zij** zijn er al. |

If you are absolutely sure that the speaker has *made a slip of the tongue*, use ***u** (see C7b).

D2. **Only use words from the CGN lexicon.** In principle this lexicon should include all Standard Dutch words in the *written* language. Here we mention explicitly a number of forms which *are* in the CGN lexicon (although it could possibly be argued that they are not part of Standard Dutch). It is compulsory to use these forms if the user does.

D2a. The following **abbreviated forms** (’k,’t,’s, ’m, ’r, ’ns, d’r, m’n, z’n, da’s, zo’n, ie) are included in the lexicon, and should be written whenever they are used.

| *you hear* | *so you write* | *i.e. you do **NOT** write* |
|---|---|---|
| /**km**oet hier weg/<br>/**tis** waar/<br>/**sa**vonds/ | **’k** moet hier weg.<br>**’t** is waar.<br>**’s** avonds. | **ik** moet hier weg.<br>**het** is waar.<br>**des** avonds. |
| /ik hepə**m** gezien/[6]<br>/ik hepə**r** gezien/<br>/kijk ə**s** hier/ of /kijk ə**ns** hier/<br>/**dər** is geen plaats/<br>/**dər**tussen, **dər**uit/<br>/**mən** huis, **zən** auto/ | ik heb **’m** gezien.<br>ik heb **’r** gezien.<br>kijk **’ns** hier.<br>**d’r** is geen plaats.<br>**d’r**tussen, **d’r**uit.<br>**m’n** huis, **z’n** auto. | ik heb **hem** gezien.<br>ik heb **haar** gezien.<br>kijk **eens** hier.<br>**er**/**daar** is geen plaats.<br>**er**tussen, **er**uit.<br>**mijn** huis, **zijn** auto. |
| /**das** nie waar/<br>/**zoon** tas heb ik ook/ | **da’s** niet waar.<br>**zo’n** tas heb ik ook. | **dat is** niet waar.<br>**zo een** tas heb ik ook. |
| /daar is **ie**/ | daar is **ie.** | daar is **hij.** |

Only use these forms where they are appropriate and **not** in the following example:

| *you hear* | *so you write* | *i.e. you do **NOT** write* |
|---|---|---|
| /**zo een** tas heb ik ook/ | **zo een** tas heb ik ook. | **zo’n** tas heb ik ook. |

D2b. The frequently used Southern Dutch forms[7] of “**ge**” and “**gij**” (and all associated verb forms) are accepted in the CGN lexicon. Do **not** change them into “je” or “jij”.

| *you hear and write* | *you do **NOT** write* |
|---|---|
| **gij** loopt snel zeg.<br>**gij moogt** weggaan.<br>**zijt ge** weg? | **jij** loopt snel zeg.<br>**jij mag** weggaan.<br>**ben je** weg? |

[6] The ə symbol stands for the schwa (the unstressed “e”).

[7] All typical Southern Dutch phenomena are marked with a grey background.

D2c. The **diminutives ending in -ke, -ske, -eke** are permitted. It is possible that some of these forms are not yet included in the CGN lexicon.
In this case you should mark them with an "***n**".

| *you hear and write* | *you do **NOT** write* |
|---|---|
| meis**ke**, tikker**ke.** | meis**je**, tikker**tje.** |
| ding**ske**, boek**ske.** | ding**etje**, boek**je.** |
| huiz**eke**, kapell**eke.** | huis**je**, kapell**etje.** |

D2d. The following **pronunciation variants,** where the "d" is abbreviated or has disappeared, are included in the CGN lexicon because they are **sufficiently *accepted* in the written language.** These and other such cases should be written where appropriate.

| *you hear and write* | *you do **NOT** write* |
|---|---|
| ik **hou** van jou.<br>**Snij** nu een ui fijn.<br>ik **rij** met de wagen. | ik hou**d** van jou.<br>snij**d** nu een ui fijn.<br>ik rij**d** met de wagen. |
| Een **ouwe** man.<br>dat gaat je niet in de **kouwe** kleren zitten.<br>Een **goeie** grap.<br>Een **dooie** man.<br>Een **kwaaie** kerel.<br>**rooie** haren. | een ou**d**e man.<br>dat gaat je niet in kou**d**e kleren zitten.<br>een goe**d**e grap.<br>een do**d**e man.<br>een kwa**d**e kerel.<br>ro**d**e haren. |

**Note:** It is **not** the intention to include **all** pronunciations variants in the CGN lexicon. Some variants are **insufficiently *accepted* in the written language**. If you hear such variants, write the Standard Dutch form.

| *you hear* | *but you **still** write* |
|---|---|
| /'t is lang gele**je**/<br>/ik ben niet echt tevre**jen**/ | 't is lang gele**den**.<br>ik ben niet echt tevre**den**. |

D3. **Do not attempt to record how words are pronounced.** After all, this would quickly lead to a sort of phonetic transcription[8]. This is **not** the idea behind an **orthographic transcription**; this may only contain words from the CGN lexicon (with the exception of *v, *d, *n, *t, *z, *x, *a, *u, or where a capital letter is written).

D3a. Do **not** mark **linking sounds** between two words. (This does not happen in the conventional spelling either).

| *you hear* | *but you **still** write* |
|---|---|
| /het leuke**n**is/<br>/speelde**n**ij goed?/<br>/liep**t**ie snel?/<br>/wadis**t**er hier aan de hand?/ | *het leuke is.*<br>speelde hij goed?<br>liep ie snel?<br>wat is er hier aan de hand? |



[8] For part of the sound fragments a proper phonetic transcription (with phonetic symbols) will be made in addition to the orthographic transcription. However the two transcriptions remain completely separate.

9

D3b. Do **not** mark where sounds are **omitted** or **not pronounced in full** (unless this is clearly due to the influence of dialect, for example diplom***d**, see D4d).

| *you hear* | *but you **still** write* |
|---|---|
| /ik ben no nie klaar/ | ik ben no**g** nie**t** klaar. |
| /da vin ik nie/ | da**t** vin**d** ik nie**t**. |
| /'k moe ier weg/ | 'k moe**t** **h**ier weg. |
| /wa is er nie goe?/ | wa**t** is er nie**t** goe**d**? |
| /**jeb**t niets gezegd/ | **je heb**t niets gezegd. |
| /ik eb **dan**twoorden/ | ik **h**eb **de an**twoorden. |
| /'t is **to**pen dat…/ | 't is **te ho**pen dat… |

D3c. Do **not** mark **small sound changes**. (Compare this with rule D5, which covers strong, dialect sounds changes, for example /bruudruuster/ written as broodrooster***z**).

| *you hear* | *but you **still** write* |
|---|---|
| /**oo**too/ (Nl) of /**o**ttoo/ (Vl) | **au**to. |
| /terap**ui**t/ (Nl) of /terap**eu**t/ (Vl) | therap**eu**t. |
| /tr**e**m/ (Nl) of /tr**a**m/ (Vl) | tr**a**m. |
| /ma**rt**/, /**sr**ift/, /**as**jeblie**f**/ | ma**rkt**, **schr**ift, **als**jeblie**ft.** |

D4. **Mark all dialect words and constructions which do not exist in Standard Dutch but do occur in dialect with *d.** Here is a summary of a few possible exceptions:

D4a. The typical Southern Dutch conjugation of articles, pronouns, adjectives and nouns :

- **ne*d** cola, **nen*d** auto, **den*d** boom.
- zijn**e*d** cola, mijn**en*d** auto, Jan, mijn**e*d m**an.
- **nen*d** blauwe**n*d** auto, **nen*d** dikke**n*d** boom, mijn**en*d** goeie**n*d** vriend.
- die**nen*d** dikke**n*d**, die**jen*d** zwarte**n*d**.

D4b. The typical Southern Dutch combination /**əkik**/. Mark this with **'k*d ik**. The form **'k*d** only occurs in the dialectic construction **'k*d ik** and can be pronounced as /**ək**/ and as /**k**/. This form should not be confused with the form **'k which occurs** in the standard written language and is pronounced as /**k**/. (see D.2.a.)

| *you hear* | *So you write* |
|---|---|
| /alzə**kik** weg moet/ | als **'k*d ik** weg moet. |
| /da weetə**kik** niet/ | dat weet **'k*d ik** niet. |
| /dat is de informatie **diekik** heb/ | dat is de informatie **die 'k*d ik** heb. |
| /ik denk **dakkik** weg ben/ | ik denk **dat 'k*d ik** weg ben. |

D4c. The forms /**de**/ and /**me**/. Consider these as dialect forms of **ge a**nd **we**. Mark them as **de*d** and **me*d**, and use the verb form belonging to **ge** and **we**.

| *you hear* | *so you write* |
|---|---|
| /**waardegij** dat?/ | waart **de*d** gij dat? |
| /**hebde** geen tijd?/ | hebt **de*d** geen tijd? |
| /nu **moede** gaan slapen. da **weettegij** wel/ | nu moet **de*d** gaan slapen. dat weet **de*d** gij wel. |
| /ik denk **damme** toffe jongens zijn/ | ik denk dat **me*d** toffe jongens zijn. |
| /**gamewij** nu al weg?/ | gaan **me*d** wij nu al weg? |

D4d. Conversational words or constructions with a hint of dialect which deviate too much from the written language should be replaced by the Standard Dutch form.

D4d. Conversational words or constructions with a hint of dialect which deviate too much from the written language should be replaced by the Standard Dutch form.

| |
|---|
| • **wulle*d, wulder*d**, **wijle*d** (= wij); **gulle*d**, **gulder*d**, **gijle*d** (= jullie); **zulder*d**, **zulle*d**, **zijle*d** (= zij). |
| • in **dees*d** geval, **dees*d** doos (= deze).<br>• een mand**e*d**, een doz**e*d**, een schijv**e*d**.<br>• deze morge**nd*d**, een dipl**om*d**, 'tzelf**st**e**d**. |
| • ze zaten daar allemaal **tope*d** (= tesamen, bij elkaar).<br>• /tis tope datniejen regent/ → 't is te hopen dat 't niet **en*d** regent.<br>• /die-en dag emmewij geen tijd/ → **dienen*d** (evt. **dieën*d**/**diejen*d**) dag **he*d me*d** (evt. **emme*d**) wij geen tijd.<br>• /kénnukkik genen tiet jong/ → **'k hen*d 'k*d** ik (evt. **kennukik*d**) **genen*d tijd*z** jong. |

D4e. Genuine dialect words (whether spoken in dialect or not):

| |
|---|
| • The Netherlands: keuje***d** (= varken); trulen***d** (= rollen); verrampeneren***d** (= ruïneren).<br>• Flanders: kortewagen***d** (= kruiwagen); savatten***d** (= sloffen, pantoffels); persienne***d** (= rolluik). |

D5. Mark words which **do belong to Standard Dutch but are pronounced with a strong dialect**, with ***z**.

| *you hear* | *you write* |
|---|---|
| /kap n**âh**/<br>/l**aa**tsepl**aa**n/ | kap nou***z**.<br>Leidseplein***z**. |
| /br**uu**dr**uu**ster/<br>/ks**aaj**n terug th**oa**s/<br>/in m**ie**n t**ie**d waz dad **alsan** zo/ | broodrooster***z**.<br>'k zijn***z** terug thuis***z**.<br>in mijn***z** tijd***z** was dat alsan***d** zo. (ter info: alsan*d = altijd) |

# E. The use of punctuation

When transcribing, use only full stops, question marks and ellipsis ( **.** / **?** / **…** ). No other punctuation marks are used! Insert these punctuation marks according to the rules and usage in written language: where would you put a full stop or question mark if you had *written* the expression yourself?

E1. Use a **full stop** at the end of an utterance. Avoid making unnecessarily long sentences: if you *could* intuitively place a full stop, then do so. Speakers often do not pause before they begin a new utterance or do pause where it is grammatically illogical. In such cases, feel free to insert a full stop where this would occur in the written language, and where your intuition tells you *can.*

- ik ga 'ns naar de Kamerleden kijken of er op dit moment een brandende vraag is. anders vraag ik de organisaties weer op elkaar te reageren.

E2. Use a **question mark** to indicate a question. Questions can be introduced by an interrogative (who, which, where, etc.) or by a verb at the start of the sentence.

- wie moet er nog koffie?
- en wie moet alles weer opruimen?
- komt Jan vandaag ook?

Sometimes only the speaker's intonation indicates that it is a question:

- Jan komt ook?
- nog koffie?

E3. Use **ellipsis** when a sentence, for whatever reason, is not finished.

- mmm niet als het… ik zal zien hoe mijn stages meevallen.
- *spreker A:* ik ben woensdag naar…
  *spreker B:* ja ja ik weet het al. je bent naar Amsterdam geweest.

Only use it to end an (incomplete) sentence or utterance. **Never** use it at the beginning or in the middle of a sentence **to mark a pause**.

- ik denk dat het waar is. (even where there is a long pause, for example between *"denk"* and *"dat",* do **not** use ellipsis)

When a sentence is broken off and then resumed, **only** insert an ellipsis where the subsequent sentence is **completely** resumed and can therefore stand on its own.

- ik ben niet goed... ik ben niet goed wakker. (*"ik ben niet goed wakker."* can stand on its own)
- ik ben niet goed uit*a... ik ben niet goed wakker. (*"ik ben niet goed wakker."* can stand on its own)
- ik ben niet goed uitgesl*a niet goed wakker. (*"niet goed wakker."* can **not** stand on its own)
- ik ben niet goed niet goed wakker. (*"niet goed wakker."* can **not** stand on its own)

E4. Make sure that **every** utterance always ends with one of the three punctuation marks.

- akkoord.
- Jan. telefoon.
- een biertje graag.
- nog koffie?
- leuke grap zeg. ggg.
- niet als het… ik zal zien hoe mijn stages meevallen

# F. Additional rules and tips

F1. Always write **numbers** in full **as they are pronounced**. The numbers from 0 to 99, the 100s, the 1000s and the 100.000s are written as one word. Everything else is written with spaces. (To reduce typing, you can write 0 to 99, the 100s, the 1000s and the 100.000s as digits. They will be converted automatically to the word form.)

| *you hear and write* | *OR you write* |
|---|---|
| drieënvijftig, **een**entwintig.<br>tweehonderd drie.<br>tweehonderd **en** drie.<br>achttienhonderd zevenendertig.<br>duizend achthonderd zevenendertig.<br>twaalfduizend tweehonderd drieënvijftig.<br>vijfhonderdduizend.<br>Drie miljoen vijfhonderd tweeëndertigduizend.<br>vijftien komma zesentwintig twaalf. | 53, 21.<br>200 3.<br>200 **en** 3.<br>1800 37.<br>1000 800 37.<br>12000 200 53.<br>500000.<br>3 miljoen 500 32000.<br>15 komma 26 12. |

If you hear the word / één /, write "één" (or "1"), if you hear /ən/ or /n/ , write "een" (without the accents).

| *you hear* | *you write* |
|---|---|
| /ən auto/<br>/één auto/<br>/er is één en ander gaande/ | een auto.<br>één auto. (ofwel: 1 auto.)<br>er is één en ander gaande. |

Write ordinal numbers and numbers which form part of a word, in full.

| *you hear and write* | *you do **NOT** write* |
|---|---|
| /de tweede keer/<br>/een veertienjarige/ | de **2de** keer; de **2de** keer.<br>een **14**-jarige. |

Unless they are part of a name (see also F4).

| *you hear* | *you write* |
|---|---|
| /de politieke partij dee zesenzestig/<br>/de snelweg aa drieënzeventig/ | de politieke partij D66.<br>de snelweg A73. |

F2. Use the **hyphen** as in written language. For example:

- een **in-**en uitgang; meelopen of **-rennen**;
- een **zee-engte**; de **West-Europese** economie;
- een **glas-in-lood**raam**; vraag-en-aanbod**situatie;
- **'s-Gravenhage**.

Many compound nouns are written as a single word, i.e. without hyphens (see also F5 and C4).

Do **not** use a dash between two words.

| *you write* | *you do **NOT** write* |
|---|---|
| de stand was 1 2 OF de stand was één twee.<br>de match Brugge Anderlecht | de stand was 1 - 2.<br>de match Brugge - Anderlecht. |

F3. Write **spelt letters (or sounds)** with capital letters as in the table below. Several spelt letters after one another are separated by a space.

| *you hear* | *you write* | *you hear* | *you write* | *you hear* | *you write* | *you hear* | *you write* | *you hear* | *you write* |
|---|---|---|---|---|---|---|---|---|---|
| /aa/ | A | /gee/ | G | /em/ | M | /es/ | S | /zet/ | Z |
| /bee/ | B | /haa/ | H | /en/ | N | /tee/ | T | /au/ | AU **of** OU |
| /see/ | C | /ie/ | I | /oo/ | O | /uu/ | U | /ij/ | IJ **of** EI |
| /dee/ | D | /jee/ | J | /pee/ | P | /vee/ | V | /eu/ | EU |
| /ee/ | E | /kaa/ | K | /kuu/ | Q | /wee/ | W | /oe/ | OE |
| /ef/ | F | /el/ | L | /er/ | R | /iks/, /ieks/ | X | /ui/ | UI |

**Never** write '**Y**', but write what you hear: **i-grec**, **IJ**, **ypsilon**, **Griekse IJ**, etc.

- yoghurt wordt gespeld als **i-grec O G H U R T**.
- **X IJ Z** zijn de drie laatste letters van het alfabet.
- hij is een lijder met lange **IJ**.
- baan met twee **A's.**
- mevrouw d'Ancona met een klein **D'tje**
- meneer **AM** De Groot.

Where letters are spelt or pronounced differently, use ***u**.

- op de lagere school spelt men laf als **le*u a*u fe*u**.
- ik heb hopen hout. met de **hasj*u** van hallo. (Vl.)

F4. Write **abbreviations** where they are used by the speaker. Write the letters in the abbreviation without spaces. Always write all abbreviations with capital letters, even where this would not (always) be the case in the conventional spelling.

| *you hear* | *you write* |
|---|---|
| /bee tee wee/, /el pee gee/, /gee es em/<br>/ex wee weeër/, /ee haa bee oojer/ of /ee haa bee ooër/<br>/bee tee wee heffing/, /gee es em gesprek/<br>de /wee see/, een /tee vee programma/<br>/aa uu bee/, /tee zet tee/ | BTW, LPG, GSM.<br>ex-WW'er, EHBO'er.<br>BTW-heffing, GSM-gesprek.<br>de WC, een TV-programma.<br>AUB, TZT (dus **niet** "a.u.b.", "t.z.t." zoals in de schrijftaal). |
| /uva/ **OF** /uu vee aa/<br>/er uu gee/ **OF** /rug/<br>/vee bee oo mavo klas/ | UVA<br>RUG<br>VBO-MAVO-klas |
| de groep /innekses/, de partij /dee zesenzestig/ | de groep **INXS**, de partij **D66** |

**Never** write an abbreviation where **none** has been said:

| *you hear and write* | *you do NOT write* |
|---|---|
| Dat wil zeggen.<br>eventueel.<br>met andere woorden . | d.w.z.<br>evt.<br>m.a.w. |

Write **acronyms** as you would in written language, but always in full with capital letters.

- een experiment van de **NASA**.
- een uitzending van de **TROS** of van de **VARA**.

If an abbreviation or acronym do not yet seem to be in the CGN lexicon add ***n** to the abbreviation or acronym.

F5. Try to follow the rules governing the writing of compound nouns as a **single word**. When in doubt, write compounds as a single word, the spelling checker will automatically correct mistakes.

| *you are not sure between* | | *you write* | *after spell checking this becomes* |
|---|---|---|---|
| woensdag morgen.<br>Hij gaat er van uit. | woensdagmorgen.<br>hij gaat ervanuit. | woensdagmorgen.<br>hij gaat ervanuit. | woensdagmorgen.<br>hij gaat ervan uit. |

In some cases there could be a difference in meaning between the compound as a single word and the separate words. This needs extra care:

| *one word* | *separate words* |
|---|---|
| dat zei ik toch **nietwaar**?<br>Hij is **tenslotte** net vader geworden (= per slot)<br>Ze is nu **tenminste** tevreden (= in ieder geval). | dat is **niet waar**.<br>dan heb ik **ten slotte** nog een opmerking (= tot slot).<br>ik wil **ten minste** 90 jaar worden (= op z'n minst). |

A great number of words are written as a single word:

- ervandaan, erbovenop, ernaartoe, ermiddenin.
- d'rvandaan, d'rbovenop, d'rnaartoe, d'rmiddenin.
- hij gaat ervan uit dat; die ziet eruit als; het zit erop.
- vijfprocentseis, lagelonenland, hogesnelheidslijn.

# G. Index

The second column gives the chapter headings; the numbers in the last column refer to the page numbers.


| | | |
|---|---|---|
| *a | C6 | 4 |
| *d | D4 | 10,11 |
| *n | C4 | 4 |
| *t | C5 | 4 |
| *u | C7, F3 | 5, 14 |
| *v | C3 | 3 |
| *x | C8 | 5 |
| *z | D5 | 11 |
| Abbreviations | F4 | 14 |
| Acronyms | F4 | 14 |
| Background sounds | C12 | 6 |
| Broken off words | C6 | 4 |
| Capital letters | C1, C2, F3, F4 | 3, 14 |
| Clearing your throat | C11 | 6 |
| Comments | C13 | 7 |
| Compounds | B1, F2, F5 | 2, 14, 16 |
| Compounds: as one word, separate or hyphenated | B1, F2, F5 | 2, 14, 16 |
| Coughing | C11 | 6 |
| da's | D2a | 8 |
| Dialect | D4, D5 | 10,11 |
| Dialectwords | D4 | 10,11 |
| Diminutives with -ke, -ske, -eke | D2c | 9 |
| Distortions | C7a | 5 |
| d'r | D2a | 8 |
| d'rtussen | D2a | 8 |
| Foreign words | C3 | 3 |
| ge, gij | D2b | 8 |
| Ggg | C11 | 6 |
| Grammatical errors | D1 | 8 |
| Hyphens | F2 | 13 |
| ie | D2a | 8 |
| Incompletely spoken words | C6, D3c | 4, 10 |
| Incomprehensible names | C10 | 5 |
| Incomprehensible words | C8, C9, C10 | 5 |
| Inflections, Zuid-Nederlandse | D4a | 10 |
| Interjections | C5 | 4 |
| 'k | D2a | 8 |
| 'k*d ik | D4b | 10 |
| Laughter | C11 | 6 |
| Linking sounds | D3a | 9 |
| 'm | D2a | 8 |
| Mispronunciations | C7b, C7c | 5 |

| | | |
|---|---|---|
| m'n | D2a | 8 |
| Names | C1 | 3 |
| New spelling | B1 | 2 |
| New words | C4 | 4 |
| 'ns | D2a | 8 |
| Numbers | F1 | 13 |
| Numerals | F1 | 13 |
| One word or not | F5 | 15 |
| Onomatopoeia | C7a, F3 | 5, 14 |
| Ordinals | F1 | 13 |
| Principle numerals | F1 | 13 |
| Pronunciation | C7, D2c, D3 | 5, 9,10 |
| Proper names | C1 | 3 |
| 'r | D2a | 8 |
| Reduced forms | D2a, D2c, D2d, D4b, D4c | 8,9,10 |
| 's | D2a | 8 |
| Sound changes, small | D3c | 10 |
| Speaker sounds | C11 | 6 |
| Spelt letters | F3 | 14 |
| 't | D2a | 8 |
| Titles of records, books, etc. | C2 | 3 |
| uh | C5 | 4 |
| uh, words interrupted by | C5, C7b | 4, 5 |
| Unfinished words | C6, D3c | 4, 10 |
| xxx | C9 | 5 |
| Xxx | C10 | 5 |
| z'n | D2a | 8 |
| zo'n | D2a | 8 |

# H. Additions

(February 2004)

## H1. Foreign words (ref. C3).

The first example in the second point under C3 <de band **New*v Kids*v On*v The*v Block*v**.> is an incorrect example, as it is the proper name of a band and should therefore **not** be given a code.

To clarify the different approaches needed for foreign word titles and proper names, the examples <de hit **Step*v By*v Step***v van de band **New Kids On The Block**> could be included after the third point under C3.

Set Latin expressions or Italian music terms are understood as Dutch words established in Dutch. They are therefore not marked with *v. If these "established words" are not yet in the lexicon, then they should be marked with an *n.

- ik heb **in extremis** de trein nog gehaald.
- hij speelde een **adagio** op de piano.
- de gemachtigde ambtenaren zouden een **status questionis*n** opmaken.

## H2. Interjections (ref. C5).

If someone says <de ramen> and extends the schwa in the word <de>, so that it actually begins to resemble the interjection <uh>, we simply transcribe <de> and **not** <d-uh*u>, <de-uh*u> or <de uh>.

## H3. Cut off words and slips of the tongue (ref. C6 and C7).

The code *a can also be used for cut off proper names, also when these are subsequently repeated in full. As an example we include <Antwer*a Antwerpen>. So instead of <Antwer-Antwerpen*u> use < Antwer*a Antwerpen >.

- de Universiteit **Antwer*a** Antwerpen is de vragende partij.
- Het is de **Belgi*a** de Vlaamse overheid die daar moet op toezien.

We also use *a for slips of the tongue which are cut short, whether or not these are followed by a resumption; i.e. : *a has priority over *u

- ze keken door het **vren*a** raam.
- ze keken door het **vren*a** venster.
- het is daar **zer*a** erg rustig.

The last example could be a case of the mispronunciation of the word <erg>, <zeer> or perhaps <zeker>. Because it is unclear which word is meant, we opt for *a in such cases. In the opposite of this example, in the sentence <het is daar zerg*u erg rustig>, we do opt for *u, because this is a case of a mispronunciation of the whole word (‘erg’ or ‘zeer’).

As C7c states we use *u in all cases of slips of the tongue and interruptions **within a word**, where the "correct" word is **not** then subsequently spoken in full. The last example, **<**ze heeft haar ring **ver-ggg-verpatst*u**>, should be interpreted exactly as it stands: if a whole word is spoken laughingly, it is not transcribed. This convention is only used if the word is **interrupted** by laughter.

- hij is nog niet naar de kapper **gewee-weest*u**.
- de verzetsstrijders van het **Kosovo-bevrijdingslever-leger*u**.
- de **jeugdlied-liefdes*u** van de schrijver.
- dat is de **mogelijkhee-heid*u**.
- heb jij dat spul nog over het gras **uitges-ge-uh-gesproeid*u**?
- dat is een **ongeluks-uh-baby*u**.
- ze heeft haar ring **ver-uh-uh-patst*u**.
- ze heeft haar ring **ver-ggg-patst*u**.

## H4. How to write down spoken language (ref. D)

Rule D1 states:

**Write down literally what is said.** This means that you should change nothing of what is said; even when the speaker uses **incorrect sentence constructions** or makes **grammatical mistakes**:

| *you hear and write* | *i.e. you do **NOT** correct this to* |
|---|---|
| de koe **dat** graast.<br>**de** open venster.<br>**hun** zijn er al. | de koe **die** graast.<br>**het** open venster.<br>**zij** zijn er al. |

If you are absolutely sure that it is a case of a *slip of the tongue*, use ***u** (see C7b).

In practice, we come across sentences such as <ik **heb** ben daar niet zo voor te vinden>. In this case <heb> is clearly a slip of the tongue, which is corrected by the speaker to <ben>. However, under rule D1 <heb> is not marked as a slip of the tongue. The last comment <If you are absolutely sure...> has definitely caused confusion.

## H5. Hyphens (ref. F2–F5);

## By Raffaëla Vlot 27-03-2001.

**General:**

WRITE DUTCH COMPOUNDS AS A SINGLE WORD

*computertafel, rekenmachine, koffiezetapparaat, loonkostenbeheersing, stoeltjeslift* etc.

**When then do you use a hyphen?**

- in compounds with vowel clashes between the parts (Otherwise the vowels would form a single sound):

| | | | | |
|---|---|---|---|---|
| aa | ee | ie | oe | ui |
| ae | ei | ii | oi | uu |
| ai | eu | iij | oo | |
| au | | ij | ou | |
| | | iji | | |
| | | ijij | | |

i.e.: *dia-avond, zee-eend, sproei-installatie, tosti-ijzer, glij-ijzer ,radio-uitzending, menu-idee*
but not in: *massaontslag, milieueffect, naijlen, parapluantenne*

---

- in compounds with coordinate, equal parts (a *journalist-cabaratier* is both a *journalist* and a *cabaratier*):

i.e.: *chef-kok, journalist-cabaretier, minister-president, rood-wit-blauw, sociaal-democratisch, koningin-moeder, Naarden-Bussum*

EXCEPTIONS: not with verbs, i.e.: *zigzaggen, hiphoppen, pingpongen*
not with two parts the same, i.e.: *beriberi, blabla, tamtam*

---

- in compounds where the left part is the crux:

i.e.: *cola-tic, rekening-courant, Staten-Generaal, Antwerpen-Oost*

[normally the right part is the crux of the compound: a *tomatensoep* is a kind of *soup*. The opposite is the case with a *cola-tic,* which is not a kind of *tic* but a kind of *cola*]

---

- in compounds with a name as the second part, in the sense of named after:

i.e.: *voorstel-Den Uyl, commissie-Geerts*

[Note: where there is a space in the same, this should be retained. Do not insert an extra hyphen. See *voorstel-Den Uyl*]

---

- also in new compounds which include a proper name

  i.e.: *nep-Rembrandt, Nuis-optreden*

  EXCEPTIONS: not if the proper name is a geographical name (which is combined with a category (= "non-proper names"):

  i.e.: *Greenwichtijd, Schipholverkeer*

  Compounds containing proper names which are included in dictionaries (and are therefore not new compounds), usually lose their hyphen. So it is *Mariabeeld* and *Nobelprijs*.

---

- also in compounds with names of languages and compound geographical adjectives:

  i.e. *Oud-Nederlands, Nieuw-Grieks*

  And also in geographic names with (one of the) points of the compass and their derivatives:

  i.e.: *Zuid-Frankrijk, Oost-Vlaanderen, Noord-Nederland en ook Oost-Vlaming, Noord-Nederlander, Noord-Nederlandse*

  also in: *Centraal-Azië, Midden-Amerika*

  EXCEPTIONS: where the geographical name was originally written without a hyphen, then it remains so e.g. *Zuidhorn*

  Comment: linked morphemes such as *oer-* and *on* are not written with a capital letter:
  i.e.: *oer-Amerikaans, on-Engels*

---

- in compounds with an abbreviation, letter or numbers:

  i.e.: *80-jarige, kleuren-TV*, meervouds-S, PVDA-forstel, top-100, TL-buis, X-benen, S-vormig, contra-VN-beleid, @-teken*

  * In the CGN project we have agreed that all abbreviations are written with capital letters, i.e. *kleuren-TV*, even though you would normally write *kleuren-tv* .

  Comment:
  - if a number is written in full, then there is no hyphen i.e.: *tachtigjarige*. In the CGN project *tachtigjarige* is written in full.
  - in a derivation with an abbreviation there is an apostrophe instead of a hyphen i.e. *D66-leider*, but: *D66'er*.

---

- If a word consists of a group of words or if there is a group of words within the word, hyphens are used between the words in the group of words, not between the group of words and a (possible) subsequent right hand part of the compound (i.e. not between ...-*smijt* and *werk*, see below):

  i.e.: *laag-bij-de-gronds, doe-het-zelver, kruidje-roer-mij-niet, jantje-van-leiden, manusje-van-alles, een staakt-het-vuren, een blik van wat-moet-je-van-me*

  *doe-het-zelfzaak, gooi-en-smijtwerk, hink-stapsprong, mond-tot-mondreclame, nek-aan-nekrace, peper-en-zoutstel*

---

- in compounds with a personal pronoun:

  i.e.: *ik-figuur, hij-vorm, het-woord*

---

- after the following prefixes (with this meaning) (or words):

| | |
|---|---|
| *adjunct* | *adjunct-commies*<br>(in the sense of '*hulp-, plaatsvervangend'*) |
| *archi* | *archi-dom*<br>('*uitermate'*) |
| *aspirant* | *aspirant-lid*<br>('*toekomstig'*) |
| *assistent* | *assistent-arts*<br>('*hulp-*') |
| *bijna* | *bijna-botsing* |
| *co* | only in *co-counsel(er/ing), co-ouder, co-schap* (easier to read)<br>but: *coassistent* |
| *concept* | *concept-tekst*<br>(the right part of the compound indicates '*tekst'*: *~begroting, ~tekst, ~verdrag*) |
| *demi* | *demi-finale* |
| *ex* | *ex-premier*<br>('voormalig') |

| | |
|---|---|
| *interim* | *interim-manager* |
| *kandidaat* | *kandidaat-notaris* |
| *leerling* | *leerling-verpleger* |
| *loco* | *loco-burgemeester*<br>*('plaatsvervangend')* |
| *niet* | *niet-roker*<br>*('ontkennend')* |
| *non* | *non-probleem*<br>*('ontkennend')* |
| *oud* | *oud-bestuurslid*<br>*('voormalig')* |
| *pro* | *pro-westers* |
| *pseudo* | *pseudo-kroep*<br>*('lijkend op, schijnbaar')* |
| *quasi* | *quasi-technisch*<br>*('schijn…')* |
| *semi* | *semi-bungalow* |
| *sint* | *sint-juttemis, sint-janskruid* |
| *Sint* | *Sint-Nicolaas* |
| *substituut* | *substituut-griffier*<br>*('plaatstvervangend')* |
| *vice* | *vice-voorzitter* |

Comment: no hyphens if the whole is not a compound or if the prefix has a meaning different to the given meaning: *locomotief, pseudoniem, oudtante*

---

- a diaeresis is used where there is vowel clash between non-compounded words (*patiënt, ruïne*) and in derivations (*beëindiging, ongeëvenaard*). Because many foreign prefixes such as *bio, macro, micro, mini, multi, neo, socio, stereo, tele, ultra* etc.. are more or less seen as independent prefixes, they get a hyphen instead of a diaeresis where there is a vowel clash:

  i.e.: *bio-industrie, macro-economie, micro-organisme, mini-emmer, multi-etnisch, neo-expressionisme, socio-economisch, stereo-opname, tele-informatie, ultra-actief*

  but: *biobak, macrobiotisch, microkosmos, minibusje, multicultureel, neoclassicisme, sociolinguïstiek, stereotoren, telewerk, ultrageluid*

  EXCEPTIONS: the prefixes *de-, pre-, re-* do not fall under this rule and do get a diaeresis
  i.e.: *deëscalatie, preïndustrieel, reïncarnatie*

  Comment: with the prefix *a-* an *n* is added, i.e.: *a+organisch* becomes *anorganisch*, *analfabeet*

---

- with compounds of three or more parts, where two or more are foreign, then a hyphen is used between the foreign parts

  i.e.: *a-capellakoor, ad-hoccommissie, human-resourcesmanagement, in-vitrofertilisatie, on-lineverbinding* (but: "*hij is on line*")

  but:
  compounds comprising two parts of which one is foreign, are written as a single word:

  i.e.: *jazzmuziek, jockeypet, jobstudent, successtory*

  If both parts are foreign and the word is established, write as a single word:

  i.e.: *airbus, bottleneck, coffeeshop, compactdisc, freelance, parttime*

  but:
  in word combinations with an adjective and a noun: write separately if both parts could be stressed:

  i.e.: *black box, happy end, heavy metal, top secret*

---

- English compounds ending with a preposition which begins with a vowel, get a hyphen:

  i.e.: *bottom-up, drive-in, lay-out, spin-off*

  but: *topdown, playback*

There are cases where a compound would be difficult to read if there were no hyphen, e.g.:

*bas-aria, jazz-zanger i.p.v. basaria, jazzzanger*

A hyphen can also make it clear which word is meant in the following examples (the hyphen is incidentally not compulsory in such cases):

*kwarts-lagen en kwart-slagen*

In addition, the writer may insert a hyphen in compounds, in particular with very long ones, for clarification in line with the spelling rules. Where the word is also easy to read without a hyphen, or where pronunciation and meaning are clear without a hyphen, DO NOT INSERT A HYPHEN, even if it would appear to be a bit clearer!!! Otherwise we would get all kinds of possible word forms, with hyphens inserted at random in various places. i.e.:

*achtertuinpolitiek, derdewereldland, heteluchtkachel, kortebaanwedstrijd, teraardebestelling, tienrittenkaart*

(From the files)

| | | |
|---|---|---|
| *brilletjes-toestand* | is simply | *brilletjestoestand* |
| *bulk-product* | is simply | *bulkproduct* |
| *conservatorium-diploma* | is simply | *conservatoriumdiploma* |

# H6. Additions and comments concerning foreign words, proper names and titles, interjections and dialect words.
# Report of a meeting on 17-10-2001 by Richard Piepenbrock.

## 1 The problem

A number of points are considered:
1 marking foreign words (ref. C3).
2 the difference between proper names and titles (ref. C3).
3 how to handle interjections (ref. C5).
4 marking dialect words (ref. D4).

## 2 The difference between foreign and Standard Dutch words

### 2.1 The treatment of foreign single word non-proper names

Examples: *bonnetterie, bonsai, burn-out, callgirl, calvados, computer, cool, disk, entremets, date, hardware, nerd*.

These are regarded as being Standard Dutch if they occur in the three volume Van Dale Groot Woordenboek der Nederlandse Taal (13e druk, 1999). This appears to be arbitrary, but the inclusion criteria for Van Dale was that words must have occurred for three years in more than one general, i.e. not subject specific source. It would be impossible to draw up better criteria within the context of CGN.

Otherwise they get '*v' and the tag 'SPEC(vreemd)'.

EXCEPTIONS:

If foreign words which are classified as Standard Dutch occur in an obvious foreign language context, then they are marked with '*v'.

Examples:

*"Ik heb al weer een nieuwe computer moeten kopen."*
*"Man ze hebben het hier over sensational*v computer*v bargains*v."*

*"Toen was de disk volgelopen."*
*"Het is zo'n rewritable*v optical*v disk*v."*

*"My*v Way*v is oorspronkelijk een Frans chanson van Claude François."*
*"Hij kreeg vaak het verwijt van het schrijven van chansons*v maudites*v."*

Take good note of Van Dale, because the sentence *'De lowbrow internet user is vooral gebaat bij up-to-date plug-and-play hardware*' does not include a single foreign word!

Notes:

1. If a non-proper name pronounced as a foreign word happens to be a homonym of a Dutch word, it is marked with '*v', even if the spelling checker does not flag it up:
"*Ik had een raar virus*v op mijn pc." (/vaIr@s/)*
*"Dat optreden van die lui was echt super*v." (/su:p@r/* with the sound *'you'*)

2. Major slips of the tongue in foreign words (which result in a different transcription) are marked with '*u', and broken off foreign words are marked with '*a', as with Dutch words. However, both are given the POS tag 'SPEC(vreemd)'. i.e. it is not necessary to mark usual mispronunciations with '*u', as in *'sweater' /swit@r/* or *'corned /kOrnEt/ beef'*.

**2.2 The treatment of foreign language multi-word non-proper names**

Examples: *anorexia nervosa, chili con carne, crème fraîche, eau de cologne, prima donna, sine qua non, vice versa.*

These words are also treated as Standard Dutch if they occur as a complete expression in the 13th edition of the Grote Van Dale and at least one of the individual parts does not occur in Van Dale (otherwise it would simply be a case of a coincidental series of independent assimilated loanwords). These are never marked with '*v', but are marked with 'SPEC(vreemd)' because the parts cannot be tagged in line with Dutch syntax.

Notes:

1. If one or more of the words also occurs on its own in a Dutch language context they are regarded as native and are therefore tagged with an appropriate tag in line with their word class in the sentence.
2. If the expression occurs in its entirety in a clearly foreign language context, then they are marked with '*v'.

Compare:

*"Voor de smaak heb ik nog wat chili N(soort,ev,basis,zijd,stan) toegevoegd."*
*"We aten die avond chili SPEC(vreemd) con SPEC(vreemd) carne SPEC(vreemd)."*
*"Op de menukaart stond chili*v SPEC(vreemd) con*v SPEC(vreemd) carne*v SPEC(vreemd) y*v SPEC(vreemd) patatas*v SPEC(vreemd) fritas*v SPEC(vreemd)."*

*"Tegen schrale huid kan ik crème N(soort,ev,basis,zijd,stan) met goudsbloem aanbevelen."*
*"Neem crème SPEC(vreemd) fraîche SPEC(vreemd) want die kun je meekoken."*
*"Dan zeggen de Fransen natuurlijk weer dat je eigenlijk een crème*v SPEC(vreemd) fraîche*vSPEC(vreemd) épaisse*v SPEC(vreemd) moet gebruiken."*

**2.3 Foreign proper names**

Foreign language proper names are never marked as '*v'.

Titles are NOT proper names. i.e.:

1. Foreign words which form part of a title retain their' *v' (but are given a capital letter under rule C2)
2. Proper names which form part of a title are still not marked with '*v' (they are after all still proper names)

The following ‘names’ are regarded as proper names:

1. a person *(George Bush, Nick Leeson, François Mitterand, Calamity Jane, Billy The Kid)*
2. a group/(music)band/theatrical troupe/... *(New Kids On The Block, The Mama's And The Papa's, The Beatles, Procol Harum)*
3. a place (*Rio De Janeiro, Orlando, Berlin/Berlijn, Nice, Praha/Praag)*
4. a street/squre/park (*Fifth Avenue, Route Sixty Six, Parc De La Villette, Madison Square, Baker Street, Hyde Park)*
5. mountains, volcanos, geysers (*Vesuvius, Matterhorn, The Old Faithful* (geyser in the USA))
6. a tunnel/bridge/building *(Pfandertunnel, Mont Blanctunnel, Sweet Track Bridge, An-Lan Bridge, Hoover Dam, Empire State Building, La Tour D'Eiffel)*
7. a company (*General Motors, Toyota, Associated Press/AP, NATO)*
8. a (cultural) organization or institution (*New York Philharmonic (Orchestra), United Nations/UN, de Neue Nationalgalerie, Ny Carlsberg Glyptotek)*
9. a festival or event *(North Sea Jazz, Crossing Border, Axion Beach Rock)*
10. a newspaper/weekly (*Bild Am Sonntag, Le Monde, New York Times)*
11. animals etc. (fauna and flora): (*Flipper, Skippy, Donald Duck* etc)
12. an internet address *(WWW Spot Dot Com, WWW RUG Punt NL, WWW RUG AC BE, WWW BBC CO UK Slash Radio One)*

(see also the list under C1 (Dutch proper names)

The following ‘names’ are regarded as titles:

1. a film *(Girlfight*v, The*v Hunter*v, Leaving*v Las Vegas)*
2. a book *(Chocolat*v, The*v Bonesetter's*v Daughter*v, Yoga*v For*v Dummies*v, The*v Adventures*v Of*v Huckleberry Finn, Max Et*v La*v Poule*v En*v Chocolat*v)*
3. a record/song *(I*v Can't*v Stand*v Myself*v, Ein*v Deutsches*v Requiem*v, Amsterdam, New York New York)*
4. a musical *(A*v Chorus*v Line*v, Cats*v, Les*v Misérables*v, The*v Phantom*v Of*v The*v Opera*v, Annie, Evita, Saturday Night*v Fever*v)*
5. a radio or TV programme *(Tatort*v, La*v Dessous*v Des*v Cartes*v, Ein*v Fall*v Für*v Zwei*v, Buffy The*v Vampire*v Slayer*v, Siska, Il*vCommissario*v Rex) etc.*

(see also the list and examples under C2 (Dutch titles).

If you are not sure about using a *v, in particular in the case of a more exotic language, you should give one. E.g. in

d*ipues*v vienes*v delhaldea*v* (Portuguese song)
*u*v lamentu*v de*v filicone*v* (Corsican song)
*sheva*v b'rachot*v* (Yiddish song)

where it is not clear if the titles contain proper names.

However, where it is clear from the context that a specific bridge/tunnel/building/newspaper etc. is meant, do not allocate a *v :

*De Gamla Tjörnbron is in 1960 ingestort (Zweeds)*
*De Akashi Kaikyo in Japan is de brug met de hoogste pilonen ter wereld (Japans)*
*Vanaf het dakterras van de Kuningan Persada heeft men een prachtig uitzicht (Indonesisch)*

Here are some more examples:

*Chris Rea, Grace Jones' laatste hit, ...*
*de band New Kids On The Block*
*Loyola University, Mechanics Institute, Massachusetts Institute Of Technology, ...*
*Metropolitan Opera, het New York Philharmonic, ...*
*heb je La*v Vie*v En*v Rose*v al eens gehoord?*
*het liedje Josephine.*
*het boek The*v Name*v Of*v The*v Rose*v.*
*Shakespeare In*v Love*v.*
*The*v Faculty*v ook die kun je vandaag zien.*

The following also are not marked with '*v':

1. Letters in names which are pronounced in a foreign language
   *('George W (/dVbl ju:/) Bush', 'J (/dZeI/) Edgar Hoover').*
2. Abbreviations pronounced in a foreign language
   *('FBI' (/Ef bi: aI/), 'BBC' (/bi: bi: si:/, 'USA' /ju: Es eI/).*

Notes:

1. Spelt foreign language letters outside the context of proper names are marked with '*v'.
   Example: *"The letters P*v (/pi:/) SPEC(vreemd) D*v (/di:/) SPEC(vreemd) en A*v (/eI/) SPEC(vreemd) staan hier for personal*v SPEC(vreemd) digital*v SPEC(vreemd) assistant*v SPEC(vreemd). Je weet wel zo'n soort elektronische agenda."*

2. Major slips of the tongue in foreign language proper names (which result in deviations in transcription) are marked '*u'. This also applies to native proper names which may or not be pronounced as foreign on purpose, e.g. *'Jambers' /dZEmb@rs/,* as if it is an American show, or in a sentence such as: *"Door al die Hollywoodfilms moeten we onderhand zeker spreken over het dagboek van Anne*u Frank*u." /En frENk/*

Broken off foreign language proper names are marked '*a', as with Dutch words. Both do however get the POS tag *'SPEC(vreemd)'*. It is therefore not necessary to allocate '*u' in for example. *'Arkansas'*, accidentally pronounced as */ArkEns@s/.*

3. Names of the days of the week and the months are regarded within CGN as proper names, although they are spelt without a capital letter. This should mean that foreign days and months in titles are exempt from being marked as '*v' i.e. *"Hotter*v Than*v July van Stevie Wonder" of "Vivement*v Dimanche van Truffaut"*. Because this is unworkable, foreign days and months are also marked '*v'.

## 3 The difference between proper names and titles

The above overview of proper names and titles, although drawn up for foreign words, also applies to Dutch words. Here are a number of additions:

Under proper names we also include:

1. A number which forms part of a proper name *(Amstel Achttien Zeventig, Plein Vierenveertig, Cinema Tweeduizend, Registratiewet Negentien Zeventig*). However, where it is simply a case of a house number then it is written with lower case letters: *Reigerstaat vierenveertig*.
2. The name of a law, royal/ministerial decision or general order in council *(Algemene Arbeidsongeschiktheidswet, wet op de Erkende Onderwijsinstellingen, de wet Waardering Onroerende Zaken, het besluit Provinciale Opcenten Motorrijtuigenbelasting)*

Under titles we also include:

1. The name of a white paper or report, which does not have the same status as a law *(de vierde nota Waterhuishouding, de nota Ruimte Maken Ruimte Delen, de voortgangsrapportage Weer Samen Naar School)*.

All category names which preceed proper names and which further specify them are not regarded as proper names or titles and are therefore written in lower case: *het ministerie van Binnenlandse Zaken, het ministerie van Middenstand En Landbouw, de vakgroep Taal En Spraak, de afdeling Personeel En Organisatie, het landelijk bestuur Scouting Nederland, de faculteit der Letteren.* Exceptions are category names which clearly form part of the name of an organization: *Orkest Van Het Oosten, Club Brugge, de Raad Van State, de Kamer Van Koophandel.*

## 4 Interjections

The '*t' was introduced in view of the problem of determining the correct notation for interjections and the fact that they do not really play a significant syntactical role for POS tagging in a sentence. However the label '*t' includes hardly any additional information compared to the label '*n' for words not previously described. We have therefore decided to replace '*t' automatically with '*n'.

## 5 Dialect words

We find that many words marked as dialect (*d) are in fact informal local (colloquial) variants of typical Dutch words, where an often rule-based vowel shift in a consonant or vowel, or particular assimilation and deletion phenomena have occurred.

Examples:

*"as (= als) Kristel hier slaapt"*
*"hij het (= heeft) niks gezegd"*
*"hej (= heb je) nog niet naar gekeken"*
*"wat heb je d'r dan an (= aan)"*

In such cases we allocate the code '*z', whereby we then have to orthographically normalize back to the standard spelling. If is not clear which word is meant, then the code '*z' cannot be allocated and the word is marked as '*d' with a pronunciation transliteration.

According to the ANS a verb form such *'wij wouden'* is classed as informal language use under point D2d of the orthographic protocol and does not therefore need to be marked. Northern Dutch diminutive forms with -ie *('stukkie', 'meissie', 'wijffie')* are not allocated '*d' of '*z' in line with the Southern Dutch variants ending in *-ke ('stukske', 'meiske', 'wijveke')*.

If in doubt, as in the *–s* forms *'derdes', '(ja)zekers', 'veels'* en *'ieders'*, the words are marked as '*d'.

Where it is a case of words which are absolutely not part of the standard vocabulary, as with *'naber'* for *'buurman', 'kappes'* for *'kool'* or *'boks'* for *'broek'*, then the label '*d' is allocated.

For reference: these words are not in the Grote Van Dale, or are allocated the characteristic *'gew*.' for *'gewestelijk'*.

Specific Flemish words which are included in Van Dale as *'Belg.N.*' and/or have received the status ‘B’ in the CGN lexicon (i.e. 'Belgicism' according to the source RBN) are regarded as equivalent to the Northern Dutch forms and are therefore not marked as '*d'.